Cartographie de l'apprentissage artificiel et de ses algorithmes

Antoine Mazieres

CARTOGRAPHIE DE L'APPRENTISSAGE ARTIFICIEL ET DE SES ALGORITHMES

Antoine Mazières (LISIS-INRA)

Thèse de Doctorat - Mention Interdisciplinaire

Université Paris Diderot - École Doctorale Frontières du Vivant (ED 474)

- Jean-Philippe Cointet (INRA) - Directeur de thèse
- Christophe Prieur (ParisTech) - Co-directeur de thèse
- Frédéric Kaplan (EPFL) - Rapporteur
- Antoine Cornuéjols (ParisTech) - Rapporteur
- Dominique Cardon (SciencePo) - Examinateur
- Jean-Gabriel Ganascia (UPMC) - Examinateur, Président du Jury

Défendue le 18 octobre 2016 à Paris, France.

"Il y a des choses comme la signification idéologique des doigts de pieds, dont peu de gens parlent et peu devraient.

Il y a des choses comme les fonctions semi-exponentielles, dont peu de gens parlent et beaucoup devraient.

Il y a des choses comme la pression des ballons au Superbowl, ou la manière dont les hommes écartent les jambes en s'asseyant, ou les robes portées à la cérémonies des oscars, dont beaucoup de gens parlent et peu devraient.

Et puis il y a ces choses comme la seconde guerre mondiale, le réchauffement climatique, les trous noirs ou l'apprentissage artificiel, dont beaucoup de gens parlent et beaucoup devraient."

Aaronson [2]

Illustration de couverture par Bernard "Papa" Mazières.

RÉSUMÉ

L'apprentissage artificiel, ou *machine learning*, est un ensemble de méthodes permettant d'établir, à partir de données, des modèles de prise de décision, de prédiction ou de classification. L' axiome plus général qui définirait ce champ de recherche est l'ambition de s'inspirer et d'imiter la capacité humaine et animale à apprendre de l'expérience. Les récents succès de ces méthodes - souvent relayés par des médias grand publics - sont seulement révélateurs de l'attention épisodique portée à des techniques qui remontent quant à elles à une cinquantaine d'années dans le contexte de l'Intelligence Artificielle et de l'informatique, et à plusieurs siècles de traditions scientifiques en mathématique, statistique, physique.

Après avoir rendu compte de ces éléments, cette thèse s'intéresse aux différentes *épistémès*, "styles de pensée" qui rythment cette communauté, en étudiant les principaux algorithmes développés pour parvenir à la prise de décision, la prédiction ou la classification. Chacun des algorithmes est envisagé de manière historique mais aussi via les contraintes techniques et théoriques qu'il porte, et les compromis d'usages qu'il impose - par exemple entre interprétabilité et efficacité. Ces "tribus" de l'apprentissages apparaissent alors comme des tentatives relativement indépendantes de parvenir à un même objectif.

Nous envisageons par la suite l'activité de ces sous-communautés algorithmiques dans le champ académique, par l'analyse de corpus bibliographiques extraits de *Web of Science*. La détection de communautés au sein des réseaux de co-citations construits à partir de ces données nous permet de mettre en lumière les structures thématiques transversales qui innervent les différents types d'algorithmes. Nous avons ainsi pu observer comment chaque discipline scientifique se place de manière spécifique dans le paysage algorithmique de l'apprentissage et entretient ou non des relations privilégiées avec les champs propres à sa recherche fondamentale. Il apparait donc, au terme de cette analyse, qu'il est plus facile pour un auteur de se mouvoir d'une thématique à une autre, que d'une méthode d'apprentissage à une autre.

Enfin, nous nous intéressons à des terrains plus ingénieriques de la pratique de l'apprentissage avec une analyse de données issues des forums de questions-réponses *Stackexchange* et du site de compétitions en ligne *Kaggle*. On y retrouve plusieurs résultats proches de ceux observés dans le champ académique, comme les disciplines les plus représentées. De nettes différences émergent cependant quant à

la diversité et la coprésence de ces algorithmes dans les compétitions et les usages des participants.

En conclusion, nous mettons en perspective certains des éléments observés dans cette étude avec les récents débats sur la place de ces algorithmes dans les politiques publiques et discutons la question de leur nature discriminatoire.

REMERCIEMENTS

Je tiens à remercier chaleureusement :

Mes directeurs, Jean-Philippe Cointet et Christophe Prieur, pour leur soutien tout au long de cette thèse,

L'INRA et l'école doctorale Frontières du Vivant de m'avoir fait confiance en m'accordant un financement,

Les membres du jury, Antoine Cornuéjols, Frédéric Kaplan, Jean-Gabriel Ganascia et Dominique Cardon pour leurs retours constructifs sur mon travail,

Joaquin Keller et Constance de Quatrebarbes pour leurs nombreuses suggestions et corrections,

Les membres du laboratoire LISIS–INRA et les toutes les personnes m'ayant aidé, accompagné, encouragé et soutenu pendant ces trois années de thèse.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

BRIBES DE CODE

ACRONYMS

ACM Association of Computing Machinery
API Application Programming Interface
EEG Électroencéphalographie
EIAH Environnements informatiques pour l'apprentissage humain
EUA États-Unis d'Amérique
GIS Geographical Information Systems
GPS Global Positioning System
GPU Graphics Processing Units
IA Intelligence Artificielle
IEEE Institute of Electrical and Electronics Engineers
INRA Institut National de la Recherche Agronomique
IRM Imagerie par résonance magnétique
KDD Conference on Knowledge Discovery and Data mining
LISIS Laboratoire Interdisciplinaire Sciences Innovations Sociétés
MIT Massachusetts Institute of Technology
MOOC Massive Open Online Courses
NLP Natural Language Processing
NYU New York University
PLOS Public Library of Science
PNAS Proceedings of the National Academy of Sciences
QSAR Quantitative Structure–Activity Relationship
SVM Support Vector Machines
UE Union Européenne
WOS Web of Science

INTRODUCTION

"[...] AI was blind and now AI see"
Amazing GrAIce, 2012
(sic.)

En quelques années à peine, l'actualité de la recherche en Intelligence Artificielle (IA), a dénoué plusieurs des défis emblématiques de ce champ de recherche. À partir de 2012, la capacité pour un dispositif informatique de reconnaître des objets sur des images et d'en décrire la scène dans un langage familier, ou, plus récemment, celle d'une voiture à circuler de manière autonome dans les rues (cf. Figure 1) sont entrés dans le domaine du possible. Ces avancées de la recherche s'appuient notamment sur l'accès à des données massives, à davantage de puissance de calcul et l'usage d'algorithmes qui ont su tirer le meilleur parti de cette combinaison : les algorithmes d'apprentissage.

(a) Un ordinateur voit et décrit "un homme qui mange un hot-dog dans une foule" (2015 [1]).

(b) Un homme surpris entrain de dormir pendant que son véhicule circule sur l'autoroute (2016 [2]).

FIGURE 1 – Quelques récents progrès d'applications en IA.

L'apprentissage artificiel, ou *machine learning*, est un ensemble de méthodes qui permet de déceler des motifs dans des données pour reproduire et optimiser le comportement ou les décisions que ces données représentent. On peut parler de manière simplifiée, mais correcte, d'automatisation par induction, ou d'apprentissage par l'expérience. Ces méthodes sont à l'honneur depuis plusieurs années et apparaissent notamment comme un élément fondamental du succès

1. https://vimeo.com/146492001
2. https://www.youtube.com/watch?v=sXls4cdEv7c

du “big data”. On présente volontiers, dans les médias comme dans des travaux de recherches prospectives, le *machine learning* comme une révolution contemporaine ou imminente, aux effets profonds sur la structure de nos sociétés, sur la cognition et sur la productivité économique. Ainsi, a-t-on pu voir déclarer la fin de la théorie scientifique [6], parce que les machines découvrent par elles-mêmes, ou la fin de la programmation [131], parce que les machines pourraient se coder elles-mêmes. Un historien va même jusqu'à formuler la possibilité que l'apprentissage artificiel soit un des éléments fondamentaux de l'idéologie dominante du siècle à venir [44, 45]. Un des traits qui encourage l'identification de l'apprentissage à un meilleur futur est, comme pour l'IA, son analogie avec une compétence humaine, le fait d'apprendre. Ces techniques seraient en mesure d'implémenter dans la machine la capacité d'apprendre ou l'intelligence.

Un premier contrepoint à cet enthousiasme médiatique est une attitude relativiste et souvent conservatrice. Celle-ci cherche à annuler l'importance de l'IA en prenant à partie l'analogie avec l'humain pour développer ce “syndrome de l'IA” (AI *effect*) que décrivait Pamela McCorduck en 1979 :

> “Cela fait partie de l'histoire du domaine de l'IA que chaque fois que quelqu'un découvre comment faire faire à un ordinateur quelque chose [...] il y a un chœur de critique pour dire, ‘ce n'est pas de l'intelligence’.” [3]
>
> McCorduck [74]

Ainsi, selon cette critique, si un système d'IA parvient à réaliser une tâche dont on pensait qu'elle requérait de l'intelligence, cette tâche n'est plus considérée comme représentative de l'intelligence. Par association, et en tant que sous-domaine de l'IA, si une procédure d'apprentissage parvient à réaliser une action jugée comme une preuve de la capacité d'apprendre, cette action est immédiatement déchue de la capacité d'apprendre qu'elle représente. Par exemple, à partir du moment où un ordinateur a remporté une partie d'échecs contre son champion mondial en 1997, les échecs ont cessé d'être un témoin de l'intelligence. Le jeu de Go, où l'ordinateur échouait, est devenu une nouvelle référence, ou preuve mondaine, qu'une machine ne pourrait jamais réaliser l'intuition qui fait de l'homme un meilleur joueur. En 2016, notamment grâce à des méthodes d'apprentissage artificiel, un ordinateur parvient à vaincre le champion du monde de Go. À ce moment même, on a pu observer de nombreuses critiques qui retiraient au Go sa qualité de “jeu intelligent” et élisait alors de nouvelles activités comme tour d'ivoire imprenable de l'exception humaine.

3. “It's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something - play good checkers, solve simple but relatively informal problems - there was a chorus of critics to say, “that's not thinking”.”

Cette critique permet de perpétuellement réduire l'ampleur des progrès de l'IA et toujours la réduire "à ce qui n'existe pas encore" [4]. Cependant, un de ces principaux désavantages est qu'elle formule le progrès de l'IA comme un conflit avec la définition de l'humain. Chaque territoire est exclusif l'un de l'autre, et à mesure que l'IA progresse, le territoire de ce qui définit l'humain se réduit, car il devrait pouvoir le définir de manière exclusive. Cet affrontement de position aux airs de guerre de tranchées peut être considéré comme une source importante du caractère science fictionnelle et souvent polémique de l'identité médiatique de l'IA, et par extension de l'apprentissage artificiel.

À l'opposé de cette trame narrative belliqueuse ou romanesque, on trouve une vision beaucoup plus inclusive de l'apprentissage artificiel et de l'IA, souvent considérée comme acquise par des acteurs plus proches de sa réalisation ou de ses disciplines scientifiques mères comme la physique ou les mathématiques. "Bien sûr, le cerveau est une machine et un ordinateur !"[116] s'exclame le médecin-neurologue Oliver Sacks, mais cela ne s'oppose en rien au fait de devoir et pouvoir "juger et éprouver", affirme-t-il. Dans le même sens, Scott Aaronson s'étonne dans la préface de son cours d'informatique d'avoir eu à revenir sur ce point :

> "Il me semblait tellement évident que le cerveau humain n'est rien d'autre qu'une 'machine de Turing chaude et humide', et tellement bizarre que je doive gâcher un seul instant de mon cours avec une question si résolue." [5]
>
> AARONSON [1]

Ainsi, pour nombre de ses acteurs, le conflit, l'opposition ou la compétition entre les capacités d'une machine, de son intelligence, de sa capacité à apprendre, avec celles des humains, ne fait pas sens. D'une part, les processus en jeu sont de même nature (induction, déduction, traitement de l'information), et d'autre part leurs contraintes ne sont pas du tout les mêmes (mémoire, capacité de calcul, architecture, consommation de ressource). Dans ce sens, la plupart des discours courants sur l'IA et l'apprentissage s'effondre au contact de leurs acteurs qui considèrent participer aux longs efforts des traditions mathématiques, statistiques, et de celles, plus récentes de l'informatique. Ces forts postulats sur la nature de l'intelligence et de l'apprentissage se différencient aussi beaucoup des discours et prospectives enthousiastes qui contrastent avec une expertise de terrain, qui observe sur-

4. Argument attribué par Douglas Hofstadter [48] à Larry Tesler. Néanmois, il semble être une adaptation à l'IA du fameux adage de Alan Kay (Apple) sur la technologie : "Technology is anything that wasn't around when you were born"

5. "it was simply self-evident that the human brain is nothing other than a 'hot, wet Turing machine' and weird that I would even waste the class's time with such a settled question."

tout de légers changements, des efforts cumulatifs qui font souvent la une d'un journal ou le succès d'un chercheur par l'accident d'une rencontre, d'une métaphore ou d'un investissement.

Comment donc rendre compte de l'apprentissage artificiel si d'une part son appréciation est ballotée entre propagande, promesses et rejet idéologiques, et d'autre part sa qualification d'intelligence apparaît comme non pertinente du fait de son trop fort ancrage dans des traditions scientifiques centenaires ? Le choix fait dans cette thèse est de partir de ce que les acteurs de l'apprentissage nomment comme tel, puis de reconstruire les sources thématiques et les attachements disciplinaires qui en font la structure concrète, notamment dans le champ académique. En ce sens, cette approche rejoint l'intuition de Allen Powell dans son histoire de l'IA :

> "Ainsi, une histoire de l'IA dans son ensemble pourrait être écrite en termes de géographie des tâches réalisées avec succès par des systèmes d'IA" [6]
>
> NEWELL [86]

Dans le chapitre 1 on voudrait montrer comment les problèmes que l'apprentissage artificiel cherche à résoudre ont été construits différemment selon leurs traditions scientifiques et méthodologiques. Ainsi, plusieurs domaines scientifiques sont proches de celui-ci alors que d'autres nous serviront plutôt à l'y opposer. Ce sont ces attractions et répulsions disciplinaires que nous envisagerons dans un premier temps, nous permettant ainsi d'introduire un exemple de résolution de problème de classification par apprentissage, puis de situer cette démarche dans le champ plus ample de l'Intelligence Artificielle (IA) et de ses inspirations prises dans le monde vivant. Fort de la connaissance de ces contextes et relations avec différentes disciplines académiques, on peut mieux introduire les traditions scientifiques qui ont permis à l'apprentissage d'être formulé et mis en oeuvre dans la deuxième moitié du XX^{e} siècle et formuler progressivement une nouvelle culture de modélisation statistique dont s'empare en partie le récent mouvement de la science des données massives.

Dans le chapitre 2, nous adopterons un point de vue plus internaliste en cherchant à montrer comment cette ambition est partagée par différents courants de pensée aboutissant à des techniques, méthodes et algorithmes différents, ayant leur propre trajectoire. Pour ce faire, chaque section de ce chapitre accompagne le lecteur dans une description détaillée des principaux algorithmes d'apprentissage en identifiant à chaque fois leurs principes d'inférence. Ce chapitre ne prétend donc pas présenter un état de l'art exhaustif des algorithmes d'apprentissage artificiel, ni en décrire le fonctionnement fin, mais il

6. "Thus a history of AI as a whole could be written in terms of the geography of tasks successfully performed by AI systems"

voudrait plutôt fournir à un lecteur néophyte une intuition de chacune de ces *épistémès* et montrer comment elles proposent chacune une solution propre à des problèmes similaires tout en s'appuyant sur des métaphores différentes. Ainsi, chaque section accompagne la description des algorithmes concernés d'éléments historiques sur leurs formulations et leurs évolutions et parcourt plusieurs problématiques transversales comme la place faite à l'intelligibilité et l'interprétabilité des modèles statistiques produits, leur propension à l'erreur et au sur-apprentissage ou leur capacité à être implémentés de manière distribuée.

Afin d'analyser comment l'apprentissage artificiel s'inscrit et déborde les champs scientifiques traditionnels, le chapitre 3 adopte une perspective résolument empirique en étudiant systématiquement la structure et la dynamique des publications scientifiques associées. L'étude des publications scientifiques est un moyen privilégié pour observer comment se structurent les dynamiques d'émergence des domaines scientifiques et, à ce titre, peut nous permettre de saisir comment chaque communauté d'algorithme émerge et interagit avec ses domaines d'applications pour construire, ou non, la cohérence du domaine de recherche de l'apprentissage artificiel.

Le chapitre 4 vise à rendre compte des usages de l'apprentissage artificiel sans se limiter au champ académique de sa recherche et de ses applications. Il s'agit donc d'observer des comportements d'utilisateurs qui explorent les outils, comme par exemple les librairies de programmation, sans forcément être à même, ou intéressés, par la compréhension de la manière dont ils sont construits et le déroulement exact des procédures sollicitées. Ainsi, dans un premier temps, on utilise les traces des utilisateurs sur des sites de questions-réponses afin de voir dans quels contextes la mention de l'apprentissage et de ses algorithmes intervient. Dans un deuxième temps, nous nous intéressons à une plate-forme de compétitions de *machine learning* afin de pouvoir observer l'impact de la nature des compétitions sur les choix stratégiques des participants.

1

L'APPRENTISSAGE ARTIFICIEL

SOMMAIRE

L'apprentissage artificiel est aujourd'hui abordé dans des contextes intellectuels divers. Dans ce chapitre, on voudrait montrer comment les problèmes qu'il cherche à résoudre ont été construit différemment selon leurs traditions scientifiques et méthodologiques. Ainsi, plusieurs domaines scientifiques sont proches de celui-ci alors que d'autres nous servirons plutôt à l'y opposer. Ce sont ces attractions et répulsions disciplinaires que nous envisageons dans un premier temps, nous permettant ainsi d'introduire un exemple de résolution

de problème de classification par apprentissage, puis de situer cette démarche dans le champ plus ample de l'Intelligence Artificielle (IA) et de ses inspirations du monde vivant (§1.1). Fort de la connaissance de ces contextes et inspirations avec différentes disciplines académiques, on peut mieux introduire les traditions scientifiques qui ont permis à l'apprentissage d'être formulé et mis en oeuvre dans la deuxième moitié du XX^{e} siècle et formuler progressivement une nouvelle culture de modélisation statistique dont s'empare en partie le récent mouvement de la science des données massives (§1.2).

1.1 CONTEXTES INTELLECTUELS DE L'APPRENTISSAGE ARTIFICIEL

Nous possédons probablement tous une intuition de ce qu'est l'apprentissage. Cette notion est communément utilisée dans le langage courant pour désigner l'acquisition de compétences et de connaissances nouvelles. En partant de cette intuition, on peut accompagner le lecteur dans une première introduction au déroulement d'une procédure d'apprentissage et saisir ainsi quelques-uns de ses enjeux majeurs (§1.1.1). Comme pour les humains, cette notion est, dans le champ de recherche informatique, subsumée par celle d'intelligence : apprendre est une forme et un témoin de l'intelligence. Ainsi, la notion d'apprentissage artificiel et sa parenté avec l'IA restent aussi floues que pour leurs analogies humaines, mais elles permettent de définir à la négative ce qu'est l'apprentissage en exposant ce qu'il n'est pas. Aussi, loin de constituer simplement une analogie, l'inspiration issue des capacités cognitives humaines, ou, plus largement, du vivant, constitue l'hypothèse centrale de cet ensemble de méthodes (§1.1.2).

1.1.1 *Quelques éléments de compréhension de l'apprentissage artificiel*

Afin de développer une première intuition sur ce qui fait l'originalité d'une approche par apprentissage, nous proposons une narration simple, mettant en scène celle-ci dans la recherche d'hypothèses scientifiques, plus précisement dans la découverte de l'équation de l'énergie cinétique (§1.1.1.1). Cette première introduction nous permet d'explorer de manière plus détaillée un exemple concret du déroulement d'une procédure d'apprentissage, où le lecteur accompagne l'algorithme dans la formulation d'une hypothèse sur un problème de classification (§1.1.1.2). Bien que la classification soit la méthode à laquelle on fait le plus référence dans cette thèse, nous présentons rapidement la typologie des formes d'apprentissage à laquelle elle appartient (§1.1.1.3).

1.1.1.1 *Un exemple de découverte*

C'est en grande partie à VOLTAIRE [143] que l'on doit la diffusion des recherches d'Isaac Newton (1642-1727) en France, initiant ainsi un vaste débat entre philosophie Newtonienne et Cartésienne. Newton étudie à Cambridge les oeuvres alors peu tolérées de Descartes représentant le monde comme un système mécanique et géométrique de matière en mouvement. Des débuts de la physique et de l'étude des astres, il hérite de plusieurs hypothèses, notamment celles de Kepler et Hook, qui lui permettent de déduire la loi universelle de la gravitation. Si les déductions de Newton s'appuient notamment sur des hypothèses qui apparaissent éronnées aujourd'hui [1], cela ne l'empêche pas de rendre compte avec exactitude de l'énergie que possède un corps du fait de son mouvement, son *énergie cinétique*. Il s'agit de l'épisode mythologique rapporté par Voltaire, où son protagoniste, assis au pieds d'un pommier [2], voyant un de ses fruits tombé, vit son moment d'*eureka*. Bien que Newton réduise sa méthode scientifique à la simple observations des phénomènes [3], cette anecdote de la pomme - peut-être une des plus connues de l'histoire des sciences - raconte une découverte, faite avec un mélange d'observations et d'hypothèses et révélée dans un moment d'intuition. On retrouve dans beaucoup de récits cette trame narrative de la découverte scientifique et de sa nature imprévisible.

Si la nature de la résolution de problème et des dynamiques de la découverte chez l'humain est loin d'être encore éclairée, l'histoire de la pomme de Newton permet d'illustrer en quoi une approche par apprentisssage est différente. Dans cette démarche on pourrait imaginer planter un capteur dans des millions de pommes, obtenir des données sur leurs chutes et attendre d'un algorithme d'apprentissage qu'il nous livre l'équation Newtonienne, sans même savoir que le soleil ne tourne pas autour de la terre. Concrêtement, le rendu de cette expérience pourrait être un tableau avec dans ses colonnes (*variables*) les données mesurées par le capteur : la masse de la pomme, sa vitesse et son énergie cinétique. Les lignes de ce tableau seraient remplies des observations de ces 3 variables pour chacune des millions de pommes considérées. Une fois ces données disponibles, on pourrait mettre en place un algorithme qui, au vue de toutes ces observations, serait en mesure de trouver l'expression (l'équation, le modèle, la fonction) la plus à même de rendre compte de l'énergie cinétique.

1. La force répulsive prise en compte par Newton dans sa démonstration s'est avérée être une force inertielle.

2. Selon STUKELEY [129], Il buvait le thé assis à sa table de jardin.

3. "Tout ce qui n'est pas déduit des phénomènes, il faut l'appeler hypothèse ; et les hypothèses, qu'elles soient métaphysiques ou physiques, qu'elles concernent les qualités occultes ou qu'elles soient mécaniques, n'ont pas leur place dans la philosophie expérimentale" [89]

Pour donner un exemple réel de l'expérience fictive décrite à l'instant, on peut citer les travaux de Schmidt et Lipson [118] qui entreprennent de redécouvrir quelques lois fondamentales de la physique en utilisant un algorithme d'apprentissage. Les données sont issues des déplacements d'un pendule double qui sont enregistrés par un capteur de mouvement (Figure 2). Afin de découvrir les invariants de ses données, ou autrement dit, les lois qui déterminent ces mouvements, l'algorithme d'apprentissage simule un processus d'évolution, c'est à dire de mutation et de sélection, entre différents composants d'une équation mathématique, comme des opérateurs ($+$, $-$, $\times$, etc) ou des fonctions ($\sqrt{n}$, cos, log, etc). Ce qui opère la sélection progressive de cette équation, c'est sa capacité à rendre compte de l'ensemble des données. Si un élément de l'équation proposée par l'algorithme augmente le nombre d'observations dont elle peut rendre compte, alors cet élément est conservé et utilisé comme source pour les mutations des prochaines générations, sinon il est écarté.

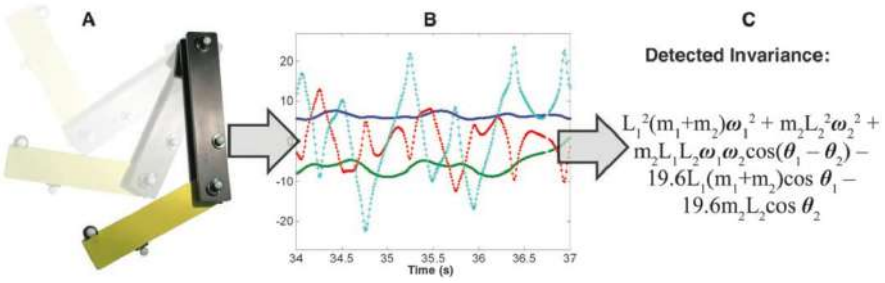


FIGURE 2 – Illustration de l'expérience de SCHMIDT et LIPSON [118].

Au final, les auteurs parviennent à redécouvrir plusieurs lois de la physique mécanique, parmi lesquelles celle de l'énergie cinétique. Cette découverte simulée illustre le propos des auteurs qu'étant donné l'accès aux outils dont ils disposaient (pendule, capteurs, etc) aucune déduction n'était nécessaire à partir d'hypothèses passées pour formuler ces lois. Ils inscrivent d'ailleurs cette méthode dans une perspective d'automatisation de la science dès les premières lignes de leur article :

> "Pendant des siècles, les scientifiques ont tenté d'identifier et de documenter les lois de l'analyse qui sous-tendent les phénomènes physique de la nature. Malgré la prévalence de la puissance du calcul informatique, le processus de trouver des lois naturelles et leurs équations correspondantes a résisté à l'automatisation. Un élément clé du défi de trouver des relations analytiques automatiquement est la définition algorithmique de ce qui fait une corrélation

observée importante et perspicace" [4]

SCHMIDT et LIPSON [118]

Le propos défendu par les auteurs de cette étude fait appel à deux notions centrales qui reviennent souvent dans les débats sur les implications du *machine learning*. Il s'agit de l'*automatisation* et de l'*induction*, ou plutot de l'automatisation par induction, qui constitue une première définition sommaire mais correct d'une approche par apprentissage. Celle-ci nous permet d'accompagner plus en détail la résolution d'un problème de classification par cette même méthode.

1.1.1.2 *Un exemple de classification*

Comme nous le verrons plus en détail dans cette thèse, il existe une grande variété de scénarios d'apprentissage. Néanmoins, le succès récent de ces approches se concentre particulièrement sur un type de tâche, la classification supervisée, l'exemple que nous détaillerons ici [5]. Par classification, on entend la dissociation ou la discrimination d'une catégorie, ou d'une classe, vis-à-vis d'une ou plusieurs autres. On peut par exemple penser à la distinction de fruits : pomme, poire et banane sont des classes de fruits, qui sont eux-mêmes une catégorie d'aliment et une catégorie de végétal. Si ce type de pensée comporte certaines failles logiques [6], il est un outil très utilisé par les humains pour décrire et décider. Toutes sortes d'organisation s'en servent pour administrer leurs activités, gérer leur stock, décrire leur ventes, etc. Une classification supervisée diffère d'une classification non-supervisée par le fait que le classement ou les résultats sont déjà connus. Ils s'agit d'éclairer les variables qui ont permis ce classement afin qu'un modèle soit à même de distinguer la catégorie d'une nouvelle observation de manière autonome. Concrêtement, cela se traduit par la présence de labels dans les données, c'est à dire une variable (*colonne*) supplémentaire que, tantot, l'on montre à l'algorithme afin qu'il s'ajuste (*entrainement*) et que, tantot, on lui cache afin de juger de son efficacité (*test*).

Dans un contexte de classification supervisée, on peut donc envisager un jeu de données comme celui montré dans le Tableau 1 pour

4. "For centuries, scientists have attempted to identify and document analytical laws that underlie physical phenomena in nature. Despite the prevalence of computing power, the process of finding natural laws and their corresponding equations has resisted automation. A key challenge to finding analytic relations automatically is defining algorithmically what makes a correlation in observed data important and insightful."

5. L'algorithme utilisé içi afin d'illustrer un exemple de classification supervisé est une régression logistique.

6. Une des plus connue étant le paradoxe de Russell sur les ensembles (un type de catégorie) : Est-ce que l'ensemble de tout les ensembles qui ne se contiennent pas se contient-il ? Si oui, alors non, si non, alors oui.

	Variables		**Label**
	Longueur sépale	*Largeur sépale*	*Espèce*
Observations (*entrainement*)	5.1	3.5	Setosa
	4.9	3.0	Setosa
	7.0	3.2	Autre
	6.4	3.2	Autre
	6.3	3.3	Autre
	5.8	2.7	Autre
	...	...	...
Observations (*test*)	4.7	3.2	? (Setosa)
	6.9	3.1	? (Autre)
	7.1	3.0	? (Autre)

Tableau 1 – Échantillon du jeu de données *Iris* [35].

dérouler notre exemple de procédure d'apprentissage. Ce tableau présente un échantillon[7] des données *Iris*, rassemblées par Ronald Fisher en 1936 [35]. Il est souvent utilisé en raison de la clareté des motifs exprimés dans les données, pour illustrer les performances d'un algorithme. Ces données de terrain traitent de classes de fleurs de type Iris et pour chacune d'elles indiquent 2 caractéristiques, ou *variables*, qui sont la longueur et la largeur du sépale. Dans notre cas, nous utiliserons ces données pour apprendre à distinguer les iris de type Setosa des autres type d'iris présents dans la base.

Comment, donc, un algorithme qui serait le même pour n'importe quel tableau du même type, peut-il apprendre de ces données et restituer un modèle permettant de prédire les classes qui nous concernent ici ? Imaginons que notre algorithme fasse face à un parterre d'Iris. Cette surface, ou plan, présente les fleurs de manière aléatoire, et leur position ne permet en aucun cas de déterminer de quelle espèce il s'agit. L'importance des données est que chacune des variables dont elle dispose pour décrire les fleurs va constituer une dimension pour les représenter dans un nouveau plan où chaque fleur apparait à une position qui la rapproche de ses pairs (Figure 3). Par dimension, on entend un nouvel axe qui définit la position du point représentant la fleur, ce qui rend difficile d'en visualiser plus de deux, mais permet d'en calculer un très grand nombre. Les observations sont donc projetées sur un nouveau plan, abstrait, dont la géographie est déterminée par les caractéristiques de chacune des fleurs. Une fois ce plan établi, l'algorithme d'apprentissage va tracer une ligne pour séparer chaque classe de toutes les autres. Si les données le permettent, c'est à dire, si elles contiennent les motifs pertinents à cette fin, la répartition abs-

7. Le jeu de données original contient 150 observations et 4 variables.

traite des fleurs dans ce plan est faite de manière à ce que le trait devant séparer une classe des autres soit le plus facile à dessiner.

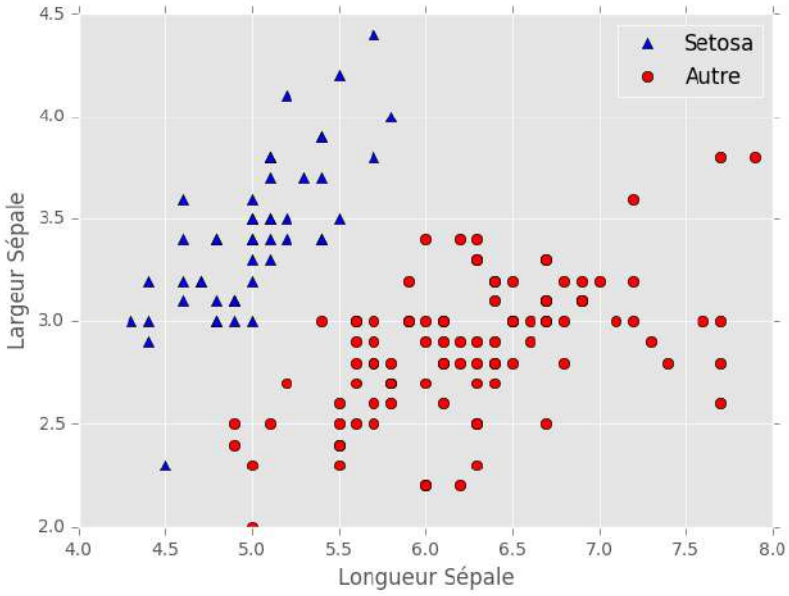


FIGURE 3 – Visualisation de la base de donnée *Iris*

Ce trait représente le modèle que produit l'algorithme. Celui-ci peut être, dans le cas de notre exemple, traduit par une équation simple. Comme on le voit dans le Tableau 2, cette équation est composée des variables de nos données, chacune associée à un poids. Ces poids constituent les paramètres du modèle qui vont subir des modifications au cours de l'apprentissage. L'établissement de ce modèle est initié par un trait aléatoire, ne tenant aucun compte de la position des points repésentant les fleurs. Ensuite, l'algorithme mesure son erreur et réajuste son modèle en fonction de celle-ci, jusqu'à ce qu'un modèle satisfaisant soit trouvé. C'est là tout l'intérêt d'une telle procédure.

L'algorithme calcule son erreur en calculant le coût de son modèle - i. e. le nombre d'éléments mal classés par celui-ci. Ce coût est une fonction dont les propriétés permettent à notre algorithme de deviner comment modifier les paramètres, à savoir les augmenter ou les diminuer, afin que le modèle ait un coût moindre à la prochaine itération. Dans l'exemple décrit ici, la fonction de coût est suffisamment simple pour que suivre le sens de sa pente (*dérivée*) nous permettent d'obtenir, pas à pas, la solution optimale à notre problème, c'est à dire une ligne droite dans le plan des données, qui sépare au mieux les points représentant les iris setosa des autres. C'est ce processus

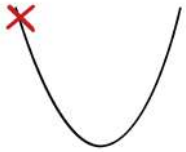
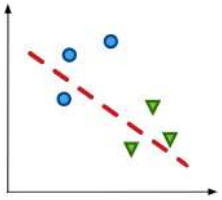
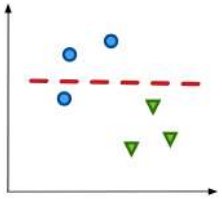
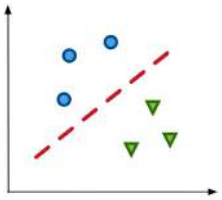
Erreur	Modèle	Équation
		$1 \times \text{longueurSépale} + 2 \times \text{largeurSépale}$
		$1 \times \text{longueurSépale} + 1 \times \text{largeurSépale}$
		$2 \times \text{longueurSépale} + 1 \times \text{largeurSépale}$

Tableau 2 – Erreurs pour l'apprentissage d'un modèle de classification des Iris Setosa.

d'itération et d'amélioration progressive qu'une métaphore vient caractériser d'*apprentissage*.

Le modèle optimal, ainsi obtenu doit ensuite être testé, ce qui explique que dans le Tableau 1 on ait séparé une partie des données de celles utilisées pour apprendre. Si la construction du modèle par l'algorithme d'apprentissage s'est déjà vu présenté les données de test dans sa phase d'apprentissage, celles-ci ne permettent plus de déterminer si le modèle est correct, et un simple apprentissage "par coeur" des cas exposés dans les données produiraient un résultat toujours juste. Ce que la phase de test cherche à déterminer, c'est la capacité de l'algorithme d'apprentissage à avoir trouvé des motifs dans les données qui permettent de reproduire une classification pertinente. On parle alors, comme pour une idée ou un concept humain, de sa capacité à *généraliser*, c'est à dire à faire sens dans des situations autres que celles qui ont permis l'apprentissage.

L'exemple présenté ici est une application simple d'un algorithme d'apprentissage très courant, la régression logistique. La régression est, en statistiques, un terme générique qui regroupe un ensemble de méthodes visant à étudier les relations d'une variable avec une ou plusieurs autres. Son origine est attribuée généralement à Francis Galton (1822-1911). Une regression est dite logistique si elle utilise la fonction *logit*, qui permet de donner des propriétés à la procédure facilitant, notamment, le calcul de l'erreur. On retrouve en 1944 des premiers usages de cette procédure dans les travaux de Berkson sur

les essais biologiques [9] et de manière plus abstraite en 1958 dans ceux de Cox [26] qui généralisent son usage à l'étude de toutes séquences comportant deux classes (*classification binaire*).

Cet algorithme simple ne recouvre pas l'ensemble des possibles parmi les procédures d'apprentissage artificiel. Au sein même des méthodes de classification supervisée, il permet une moindre performance des prédictions mais évite de nombreux problèmes et offre une simplicité procédurale qui donne beaucoup de marge de manoeuvre à qui l'implémente, pour s'adapter à son contexte d'utilisation. Cette simplicité nous a permis notamment de développer une première intuition de l'application concrete d'une procédure d'apprentissage. Mais cette intuition ne vaut que pour une procédure supervisée et il convient d'explorer deux autres familles qui porte la même ambition d'apprentissage mais avec des contraintes différentes.

1.1.1.3 *Autres types d'apprentissage*

L'exemple que nous venons de voir est celui d'une procédure de classification supervisée. C'est à dire que son objectif est de classer des éléments selon des informations connues, des labels, qui définissent strictement à quel groupe doit appartenir tel ou tel observation. À minima, il s'agit d'une approximation de la fonction qui lie les variables aux labels. Comme nous l'avons déjà indiqué, il s'agit de la méthode d'apprentissage la plus courante et celle qui constitue le coeur du succès contemporain de ces techniques. À ce titre, il s'agit de celle à laquelle nous ferons le plus référence dans cette thèse. Cependant, il est important de décrire les deux autres grandes familles d'apprentissage : l'apprentissage non-supervisé, c'est à dire sans label aucun, et l'apprentissage par renforcement, que l'on peut considérer comme un compromis entre les deux autres méthodes.

L'apprentissage non-supervisé porte l'ambition de permettre à une procédure d'analyse de découvrir les données, leurs structure et dynamique, sans être informée de ce qu'elle doit observer, sans objectifs précis à poursuivre. Dit plus simplement, on attend d'une démarche non-supervisée qu'elle apprenne sans qu'on lui dise quoi apprendre. Cette injonction peut sembler contradictoire en ce qu'on ne peut pas juger de sa réussite si aucun objectif n'est déclaré, et, si un objectif est déclaré, on peut parler, au moins indirectement, d'un type de supervision de la procédure. Malgré cela, ce champ de recherche très actif est souvent représenté comme le coeur de la question de l'apprentissage et de l'intelligence artificielle. C'est en ce sens que plusieurs auteurs affirment :

> "Nous attendons de l'apprentissage non-supervisé qu'il devienne beaucoup plus important au long terme. L'ap-

> prentissage chez l'humain et l'animal est en grande partie non-supervisé : on découvre le monde en l'observant et non en étant dicté le nom de chaque objet." [8]
>
> LeCun, Bengio et Hinton [67]

Une des principales utilisations contemporaines de l'apprentissage non-supervisé est le regroupement, ou *clustering*, dont le but est de trouver des classes pertinentes au sein d'un corpus de données qui ne comporte pas d'information sur celles-ci. Pour reprendre l'exemple du jeu de données Iris, on attendrait d'une telle procédure qu'elle distingue les espèces de fleurs sans avoir aucune indication sur celles-ci, si ce n'est les données décrivant chaque observation. Pour ce faire, les algorithmes de clustering observent la similarité entre les observations et déterminent quels regroupement permet d'en rendre le mieux compte. Géométriquement, c'est à dire dans le plan de représentation des observations et de leurs variables, la similarité s'exprime le plus souvent par la proximité. Cependant le concept de distance qui fonde celui de proximité fait l'objet de nombreux formalismes mathématiques au-delà de celui, le plus intuitif, euclidien. En effet, on retrouve plusieurs manières de mesurer la distance entre deux points par exemple les distances de Manhattan, de Levenshtein, de Chebyshev, etc, chacune mettant en lumière des informations différentes sur la distance qui sépare des points dans le plan. Cette diversité de méthodes s'exprime aussi dans la manière de regrouper les points une fois leur proximité calculée. On retrouve ainsi de nombreuses méthodes pour ce faire, parmi lesquelles KMeans, DBScan, regroupement hiérarchique, etc. La Figure 4 illustre comment ces méthodes de regroupement peuvent trouver des groupes pertinents (Figure 4a), ou trouver des groupes quand il n'y en a pas (Figure 4b), ou n'en trouver aucun lorsqu'il y en a (Figure 4c), ou ne pas trouver les bons (Figure 4d).

De plus, certains algorithmes de regroupement associent plusieurs groupes à chaque observation, certains sont probabilistes et chaque exécution de la procédure peut retourner des résultats différents. Ainsi, comme l'affirme Jain [53], "une structure de regroupement est valide si elle ne peut raisonnablement pas avoir été produite par hasard ou comme un artefact d'une procédure de regroupement" [9]. Alors que l'apprentissage supervisé a cette propriété rassurante d'optimiser l'approximation des entrées et des sorties, il est en revanche difficile de caractériser les résultats issus des techniques non-supervisées de bons ou mauvais. Elles sont à minima une procédure non-aléatoire, rendue possible à grande échelle, pour l'exploration et la formulation

8. "We expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised : we discover the structure of the world by observing it, not by being told the name of every object."

9. "A clustering structure is valid if it cannot reasonably have occurred by chance or as an artifact of a clustering algorithm"

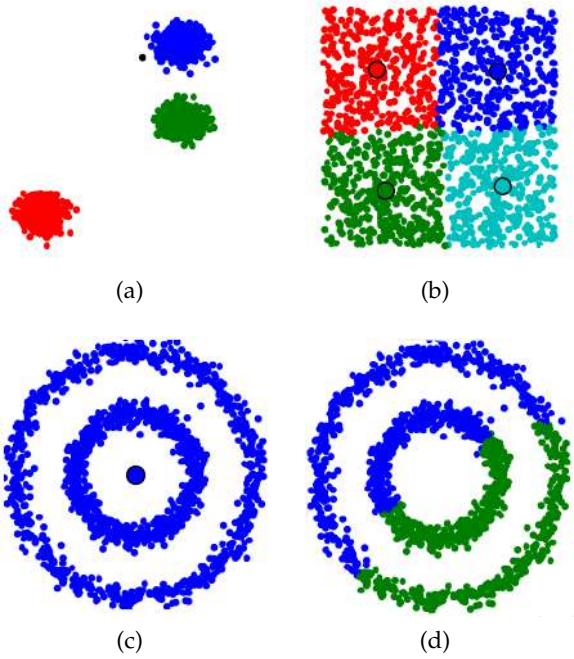


FIGURE 4 – Exemples de regroupement.

d'hypothèses. Dans ce sens, l'utilisateur peut chercher parmi toutes les méthodes qui font une procédure de regroupement, celle qui produit une structure qui fait sens par rapport à son évaluation *a priori*. Ce mouvement de va et vient entre attentes, *a prioris*, et la pertinence des données, permet à une intuition sur la structure des données de passer à l'échelle de l'ensemble d'un corpus.

À mi-chemin entre les procédures supervisées et non-supervisées, on trouve notamment l'apprentissage par renforcement. Sa description générique fait assez naturellement appel aux systèmes vivants en ce que l'objet de l'apprentissage est un comportement et la méthode est son interaction avec son environnement. L'agent de ce comportement dispose d'un ensemble d'états possibles qui lui sont accessibles afin d'effectuer un ensemble d'actions. L'agent apprend quelle suite d'actions réaliser en percevant des signaux de son environnement qui soit la "récompensent" soit le "punissent" pour l'aider à apprendre une stratégie pertinente qui est l'objet de la procédure d'apprentissage. Cette procédure est supervisée en ce que l'artisan de cette procédure doit à la fois définir les actions et états possibles et surtout les comportements qui sont valorisés ou non. Mais l'algorithme repose surtout sur des indications plus que sur un objectif à optimiser ce qui amoindrit la supervision dont il fait l'objet pour qualifier sa stratégie.

Cette rapide mention des grandes familles de l'apprentissage artificiel nous permet d'envisager l'ampleur de leurs ambitions au-delà des récentes prouesses des seules techniques supervisées. Ainsi on peut

envisager des analogies plus générales pour caractériser les objectifs de ce champs de recherche et, ainsi, le définir par rapport à d'autres champs de recherche.

1.1.2 *Éléments de définition et d'analogie de l'apprentissage artificiel*

Le récent succès des méthodes d'apprentissages a permis à de nombreuses initiatives pédagogiques de voir le jour en la matière, notamment par la publication de cours en ligne (MOOC) et de livres dont l'ambition est de diffuser ces connaissances en les destinant à un public moins acquis que celui de ses domaines d'origines (statistiques, mathématiques, physique, etc). Nombre de ces initiatives ont été étudiées dans le cadre de cette thèse, au moins en partie, et constituent ainsi un terrain pour décrire et analyser comment les acteurs de ce domaine définissent leur champ de connaissance.

La manière dont un enseignant introduit le domaine que son cours s'apprête à divulguer est révélatrice de comment, lui-même, se positionne dans sa discipline de "rattachement" qui, bien souvent, subsume son activité. Ainsi, ces "introductions" constituent un matériel pertinent pour définir l'apprentissage artificiel et comment ses acteurs et diffuseurs se situent par rapport à leurs entourages thématiques. Nous nous reposons principalement sur 3 exemples qui illustrent bien la diversité des introductions rencontrées [4, 47, 90]. Afin de les regrouper, nous parlerons des introductions par la négative qui opposent le *machine learning* à d'autres méthodes ou perspectives. Ceci permettra de donner plusieurs exemples d'approches qui ne sont pas de l'apprentissage au sein même du domaine de l'Intelligence Artificielle (§1.1.2.1). Ensuite, nous verrons un courant plus constructif de définition, qui part d'une analogie avec l'humain, ce qui sera l'occasion d'explorer une première fois les inspirations réciproques entre l'apprentissage artificiel et les sciences du vivant (§1.1.2.2).

1.1.2.1 *Des définitions par le négative de l'apprentissage automatique*

Afin d'introduire une notion, il peut parfois être très confortable de l'opposer à une autre. Si le contexte commun de ce qu'on explique et de ce à quoi on l'oppose est assez clair, cette méthode est efficace pour faire comprendre dans quelle zone "contextuelle" on se situe. C'est en faisant usage de ce type de définitions, par la négative, que nous introduisons dans cette partie l'apprentissage artificiel.

L'APPRENTISSAGE CONTRE LA PROGRAMMATION

Dans son cours en ligne, Geoff Hinton [47] commence par définir l'apprentissage artificiel en l'opposant à la pratique classique de la programmation, ou, plus précisemment, à l'acte d'écrire un programme. En effet, selon Hinton, le *machine learning* est nécessaire quand nous tentons d'écrire un programme alors qu'on ne sait même pas comment notre cerveau réalise cette opération. Imaginons un cas simple où nous cherchons à obtenir un classifieur qui distingue homme et femme, en ignorant les cas tiers pour la simplicité de l'argument. De multiples facteurs (organes sexuels notamment) traduisent l'appartenance à ces classes, et d'autres variables plus probabilistes peuvent, une fois corrélées, renforcer cette prédiction (poids, taille, masse musculaire, tour de poitrine, etc). Si l'on dispose de toutes ces données sur un échantillon d'observations, on peut facilement imaginer qu'il soit possible d'écrire un programme "classique" permettant de distinguer à coup sûr un homme d'une femme. Par programme classique on entend une suite d'instructions lisibles faite de propositions conditionnelles, du type Si $X > Y$ et $Z = 0$ alors Homme, sinon Femme, ou inversement.

Reposons ce problème, mais cette fois-ci, au lieu de disposer de données déjà extraites, nous ne disposons que d'une photo de la personne à classifier, dans un contexte (ville, rue, intérieur, campagne, etc) et une luminosité toujours changeante. Alors que nous sommes, nous, humains, capables dans la plupart des cas de classifier homme ou femme à partir d'une photo, il nous est impossible de décrire une procédure pixel par pixel qui puisse rendre compte du genre de la personne sur la photo. Même si nous étions capables de décrire comment notre cerveau traite chaque pixel pour reconstruire les formes puis considérer les éléments pertinents à la classification, cela serait beaucoup trop compliqué à implémenter, sans aucune règle simple et constante sur laquelle s'appuyer. Ce type de programme qui nécessite un grand nombre de règles faibles et changeantes, est l'objet, selon Hinton, de l'apprentissage artificiel qui déplace la question de résoudre ces problèmes à celle de dégager automatiquement des règles de résolution à partir de l'observation d'un nombre massif d'exemples.

Si cette méthode s'impose dans cet exemple par la contrainte - i. e. l'impossibilité d'user d'un autre moyen - il s'agit aussi d'une opportunité pour diminuer le besoin de programmer "à la main", même s'il s'agit de choses relativement accessibles. Ainsi, Hinton n'hésite pas à mettre en balance le prix de la puissance de calcul nécessaire à ce type d'apprentissage avec le prix que coûte un programmeur humain pour résoudre le même problème. Si une telle perspective n'est envisageable que depuis récemment, elle est déjà mentionnée dans un des articles fondateurs du *machine learning*, où l'auteur, Arthur Samuel, affirme que "programmer des ordinateurs pour apprendre

de l'expérience devrait éliminer une grande partie de l'effort de la programmation des détails." [10] [117].

Ce résultat peut sembler encore lointain, surtout lorsqu'un grand nombre de politiques de formations universitaire et secondaire pour apprendre aux étudiants à programmer est en marche, considérant cette habilité comme cruciale sur le marché du travail à venir [11]. Néanmoins une recherche récente [55] voulant démontrer la "déraisonnable efficacité" d'un algorithme d'apprentissage, a réussi à produire un code en apparence très crédible, à partir d'un entrainement sur l'ensemble du code source de Linux. Le résultat semble, de loin, parfaitement crédible (Bribe de code 1). On y trouve un respect de la syntaxe (indentation, passage à la ligne, accolade), la présence de commentaires, etc. Néanmoins n'importe quel programmeur un peu au fait du language utilisé (C) et de l'objet programmé (le kernel Linux) se rend rapidement compte que ce code informatique est complétment absurde, voir plein d'humour, et ne réalise aucune action si ce n'est de ne pas pouvoir être exécuté.

Une autre approche à ce problème a été développée par Zaremba et Sutskever [149] qui entrainent un algorithme d'apprentissage à écrire un programme capable de réaliser des additions à partir d'une série d'exemples de cette opération. Le résultat est un code très peu intuitif et trop compliqué étant donné la simplicité de la tâche mais qui parvient à réaliser l'opération. Néanmoins, il parvient à ce résultat seulement 99% du temps.

Ces premiers résultats dans le sens de l'hypothèse de Hinton, semblent laisser entrevoir que le *machine learning* pourrait ne pas être une solution seulement quand un problème est trop difficile à résoudre mais plutot quand des données permettent de le résoudre de cette manière et ainsi d'éviter le cout et l'incertitude d'une production humaine.

L'APPRENTISSAGE CONTRE LES MATHÉMATIQUES ET LE DESIGN

Le cours réalisé à l'Université de Caltech depuis 2012 sur l'apprentissage automatique a été entièrement reproduit sur un plateforme de cours en ligne [4], et a donné lieu à la publication d'un livre éponyme, *Learning from data*, qui constitue un des best-sellers du domaine [3]. Cet enseignement introduit le *machine learning* en décrivant les conditions qui justifient son utilisation :

10. "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort"

11. Par exemple, dans son discours sur l'état de la nation en 2016, le président américain a annoncé un plan fédéral (4 milliards de $) pour que tous les élèves du pays se voient offrir des cours de programmation.

Bribe de code 1 – Code informatique généré à partir d'un entrainement sur le code du noyau Linux [55]

```
/*
 * Increment the size file of the new incorrect UI_FILTER group
     information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] |
      PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

- Le fait que des données existent et soient disponibles, sans lesquelles l'expérience dont à besoin l'algorithme pour apprendre n'existe pas.
- Le fait que dans ces données existent un motif qui permettent de résoudre le problème posé. Sans cela, il n'y a rien à chercher dans les données.
- Le fait qu'on ne puisse pas mettre le doigt sur ce motif mathématiquement.

C'est bien avec cette troisième condition exposée par Abu-Mostafa que l'apprentissage prend une position méthodologique propre en s'opposant à l'approche "mathématique". Cette opposition peut être assez déroutante dans un premier temps au vu du nombre d'éléments mathématiques qui sont empruntés dans toutes les procédures d'apprentissage artificiel (analyse, algèbre linéaire, etc). On peut donc penser à plusieurs interprétations possibles et complémentaires de cette opposition. La première serait la même que celle exposée précédemment (§1.1.1.1), comparant l'approche de Newton à celle de l'apprentissage pour découvrir la loi de gravitation. Dans ce cas, on parle d'approche mathématique surtout pour décrire la dérivation et la combinaison d'une équation à une autre, en se fondant en premier lieu sur les propriétés de celles-ci. Une deuxième interprétation possible de l'approche dite "mathématique" serait d'imaginer un humain à la place de l'algorithme d'apprentissage, à la fois dans l'acquisition de l'expérience des données, et dans l'intuition et la perception des motifs qui constituent le modèle. On peut, par exemple, penser à un analyste qui perçoit plusieurs fonctions (log, racine carrée, etc) qui permettent de rendre compte des corrélations entre les variables, et parvient ainsi à construire un modèle. Enfin, l'approche "mathématique" peut avoir un sens encore plus général, de démarche *a priori*, où on établit un modèle dont la distance avec l'objet étudié est moins considéré comme une faille que comme un acte conscient de *design*, de modelage de ce que le modèle va réguler. C'est cette interprétation de l'approche "mathématique" comme *design* qui est la plus développée dans le livre issue du cours. Selon les auteurs, il s'agit d'une approche par spécifications où celles-ci viennent remplacer la nécessité des données.

Cette perspective fait écho à l'analyse de Dominique Cardon [22, 23] de différents types d'algorithmes. L'auteur de *À quoi rêvent les algorithmes* propose en effet une typologie de ceux-ci dont deux éléments recoupent en partie l'opposition établie par Abu-Mostafa. L'approche dite mathématique est invoquée par celle dite du principe d'autorité, qui se place "au-dessus" des données, forte d'une vision ou de spécifications *a priori*. Un exemple donné par Cardon est l'algorithme du *PageRank* [93] qui permet un classement des pages web en valorisant celles qui citent et sont citées par d'autres pages. Ce postulat

a priori fait autorité pour déterminer le mérite d'une page à être présentée à un utilisateur. Il y a un choix fait en amont qui récompense les pages qui "jouent le jeu" d'un web interconnecté, référencé, et inspirent donc une légitimité démocratique à leur algorithme de classement. Cette approche "au-dessus" des données fait contraste avec une approche "par-dessous" qui repose sur ce que nous pourrions appeler "la tautologie des traces". Dans le cas du classement des pages webs, les traces laissées par les utilisateurs pendant leurs navigations permettent de proposer ce que les traces prédisent : les pages que les utilisateurs visitent sont celles qu'ils sont le plus probable de visiter. Cette proposition n'impose aucune autorité ou principe *a priori*, outre le fait que ce qui émerge des comportements des utilisateurs est ce qui structurera ceux-ci.

1.1.2.2 *Analogies et inspiration avec le vivant*

Si des définitions par la négative permettent de situer l'apprentissage machine par rapport à plusieurs autres approches, trouver une définition propre à ce champ, qui le définisse en tant que tel, nécessite de revenir à la métaphore de l'apprentissage. En effet, on trouve celle-ci au coeur des premiers travaux de ce domaine de recherche, comme un élément concret et formel de sa définition.

L'APPRENTISSAGE COMME UNE APPARENCE

Plusieurs matériaux pédagogiques sur le sujet [90, 115] considèrent qu'on doit la première référence à l'apprentissage artificiel à Arthur Samuel dans son article de 1959, *Some studies in Machine Learning Using the Game of Checkers* [117]. Ce papier fondateur, expose une expérience où l'auteur réalise de nombreuses parties d'échecs avec un ordinateur qui, au fil de l'enregistrement et de l'analyse de celles-ci, parviendra à vaincre son architecte. L'apprentissage, notion centrale dans le récit de l'article, est défini comme suit dans les premières lignes de l'étude :

> "Les études rapportées ici traitent de la programmation d'un ordinateur afin qu'il se comporte de telle manière que, s'il s'agissait d'une action humaine ou animale, elle serait décrite comme un processus d'apprentissage." [12]
>
> SAMUEL [117]

Ainsi, pour Arthur Samuel, l'apprentissage machine, c'est ce qui nous semble être de l'apprentissage humain, ce que notre intuition nous

12. "The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning."

porte à considérer comme tel. En d'autres termes, si nous sommes intéressés par la capacité d'apprendre des humains afin de la transmettre aux machines, il faut commencer par faire en sorte d'avoir l'impression que l'ordinateur apprend. Cette invitation à l'imitation peut sembler fragile pour fonder une méthode scientifique, mais on peut trouver dans l'histoire des sciences et des technologies de nombreuses analogies similaires avec l'humain ou l'animal pour inspirer les premières explorations et expériences d'un projet. À titre d'exemple, les premiers efforts de l'aviation s'inspiraient volontiers d'éléments morphologiques des oiseaux pour imiter ce qui vole afin d'explorer comment voler. La Figure 5 illustre ce cas en montrant l'avion III de Clément Ader dont les hélices avaient la forme de plumes, les ailes une forme inspirée de chauves-souris et, en y regardant de plus près, couvertes de petites plumes.

FIGURE 5 – Avion III de Clément Ader exposé au Conservatoire National des Arts et Métiers à Paris.

Si les termes de cette imitation sont peu définis, d'autres définitions plus formelles viennent imposer des "garanties", c'est à dire ce qu'il faut pouvoir observer de l'algorithme pour considérer qu'il "apprend". Ainsi, Tom Mitchell [80] affirme qu'un tel programme informatique "peut être décrit comme apprenant de l'expérience E la tâche T avec une performance P, si sa performance pour T, mesurée par P, s'améliore avec l'expérience E" [13]. L'apprentissage artificiel est donc "l'étude des programmes informatiques qui s'améliorent automatiquement avec l'expérience". Mitchell file ainsi la métaphore de l'apprentissage en caractérisant les données d'expérience, faisant à nouveau référence à l'intuition du lecteur concernant un comportement humain ou animal. Chaque observation d'un jeu de données constitue une expérience vécue par l'algorithme qui lui permet de s'améliorer, d'apprendre.

13. "A computer program is said to learn from experience E with respect to some class of tasks T and performance P, it its performance at tasks in T, as measured by P, improves with experience E"

L'APPRENTISSAGE COMME UNE HYPOTHÈSE

Comme nous venons de le voir, l'analogie avec l'apprentissage n'est, à l'origine, pas simplement une métaphore mais une invitation à l'imitation et à l'évaluation par la comparaison avec l'humain et l'animal. Cependant, à mesure que la définition de ce qui est à imiter se précise en tant qu'optimisation par les données, de nombreux chercheurs déclinent l'analogie mère pour se concentrer sur les aspects mathématiques de cette procédure inductive. Cette intention se traduit surtout dans l'étude précise d'algorithme, où l'inspiration du vivant s'estompe avec la formulation de résolution de problèmes concrets pour améliorer les performances ou certaines propriétés de ces procédures. Dans ce contexte, ce qui reste de l'apprentissage est une hypothèse sur la nature de l'acquisition de nouvelles compétences. On retrouve une formulation de cette hypothèse dans le cours en ligne de Andrew Ng [90] qui constitue un des enseignements les plus suivis en la matière :

> "Le cerveau fait tellement de choses extraordinaires, il semble que pour l'imiter il faudrait obtenir un grand nombre de logiciels pour reproduire toutes ces aptitudes. Mais une hypothèse fascinante est que la manière dont un cerveau fait toutes ces choses est le fruit d'un seul et unique algorithme d'apprentissage." [14]
>
> NG [90]

En ces termes, Ng met en avant un postulat de la notion d'apprentissage, c'est à dire qu'un mécanisme unique est la source de toute acquisition de savoir ou de compétence. Pour illustrer son propos, Ng, fait appel à plusieurs expériences de neurobiologie allant dans le sens de la confirmation de ce postulat, l'une d'entre elle étant le rewiring (*rebranchement*) qu'on trouve notamment dans les travaux de ROE et al. [109]. Cette expérience atteste qu'une zone du cerveau est spécialisée dans le traitement d'un type de signal, par exemple auditif, et qu'en rebranchant cette zone à une autre source d'information, par exemple la vue, le sujet parvient à maintenir sa capacité à entendre. Selon Ng, ce que cette expérience montre c'est que le cerveau possède une propriété unique, qui ne dépend pas de sa localisation, et dont la caractéristique universelle est de pouvoir traiter l'information qui lui parvient et apprendre à partir celle-ci. L'intention de Ng est de montrer que si l'on parvenait à capturer cette compétence mère, alors toutes celles en aval seraient accessibles *via* l'exercice de cette première.

14. "The human brain does so many different amazing things, it seems that if you want to mimic the brain you have to write lot of different pieces of software. But there is this fascinating hypothesis that the way the brain does all these different things is just with a single learning algorithm."

Par ailleurs, la "quête" pour l'apprentissage automatique, un "algorithme maître" [33], est une proposition pour résoudre un problème plus large qui subsume celui de l'apprentissage : l'intelligence. On parle d'Intelligence Artificielle (IA) en informatique ou en robotique un peu de la même manière qu'on fait référence à l'apprentissage, comme une métaphore, une analogie, qui permet de comparer, d'évaluer par rapport à notre intuition de ce que l'on cherche à reproduire, à automatiser : l'intelligence humaine. Dans ces termes, les chercheurs qui s'identifient à ce domaine parlent souvent de l'axiome de "résolution" de l'IA. "Résoudre l'IA" c'est trouver une procédure ou un ensemble de procédures qui reproduisent les principales propriétés de ce que l'on appelle intelligence et qui pourrait être appliquée de manière universelle à n'importe quel problème.

Dans ce contexte, on retrouve dans *Intelligence Artificielle : Une approche moderne* [115], considéré comme un livre de référence dans plus d'un millier d'universités, une des cinq parties de l'ouvrage consacrée à l'apprentissage. L'apprentissage y est une hypothèse, une piste, pour résoudre l'IA, au même titre que la recherche de solutions parmi des arbres de possibilités, les combinaisons logiques entre des représentations de connaissance, qui font l'objet d'autres parties. Cet exemple, tiré d'un contenu pédagogoqique de référence, invite à considérer l'hypothèse de l'apprentissage que nous avons identifié, comme incluse, au moins en partie, dans les efforts pour résoudre l'IA.

1.2 ORIGINES SCIENTIFIQUES ET APPROPRIATIONS CONTEMPORAINES

L'apprentissage artificiel est aujourd'hui le centre de beaucoup d'attention des médias grand public et reçoit des investissements venant de sources multiples autant dans des perspectives industrielles que de recherche. Il ne se passe pas quinze jours sans qu'un journal généraliste ne fasse le point sur "la révolution de l'intelligence artificielle" ou n'explique les impacts du *machine learning* sur le "big data". Ce succès médiatique peut laisser croire à une émergence récente de ce domaine, qui vivrait ici et maintenant son premier franc succès incontesté, quand bien même ses principes et dénomination datent d'un demi-siècle. Un examen plus détaillé des dernières décennies montre qu'il n'en est rien et que les pistes de recherches explorées en apprentissage sont déjà toutes passées par des périodes de forte publicité et d'oubli, poussant même leurs acteurs à y faire référence en terme de saison. En effet, "l'hiver de l'IA" est une métaphore souvent utilisée pour décrire les périodes de disète d'intérêt et de financement qu'a connu ce domaine pendant la seconde moitié du XX^{e} siècle.

Mais dans une perspective historique qui tenterait d'identifier ce que la quête de l'intelligence ou de l'apprentissage vient catalyser, il est nécessaire de remonter à des sources disciplinaires plus anciennes et de montrer comment celles-ci viennent alimenter les ambitions du domaine dont on reconstitue ici la génèse. La statistique est le domaine historique communément rattaché à l'apprentissage artificiel. Ainsi, l'apprentissage peut être vu comme la rencontre entre une partie de la tradition statistique de plusieurs siècles et une partie des ambitions de la science informatique naissante du xx^{e} siècle. Après avoir rapidement situé ces deux sciences dans un premier temps (§1.2.1), nous pourrons voir comment cette rencontre favorise la naissance d'une nouvelle culture statistique qui porte en elle plusieurs éléments de ce que l'on dénomme plus récemment la "science des données" (§1.2.2).

1.2.1 *Statistique, Informatique et Matériel*

On aborde souvent la question de l'apprentissage en parlant d'algorithme, or ce terme recouvre un ensemble d'usages divers dont il peut être difficile d'identifier ceux qui nous intéressent ici. Pour cela, il convient de remettre le terme d'algorithme dans le contexte de la science informatique qui le définit au moins en partie (§1.2.1.1). De manière similaire, une rapide observation des débuts épars de la science statistique permet de souligner l'importance de choisir une définition et un courant plus précis en son sein pour pouvoir rendre compte des conditions nécessaires à l'émergence de l'apprentissage artificiel (§1.2.1.2).

1.2.1.1 *Informatique et Algorithme*

Le terme d'algorithme, dont on attribue l'origine à la latinisation du nom de l'algebriste perse Al-Khawarizmi (780 - 850), signifie à minima un suite d'instructions. En ce sens, une recette de cuisine, des indications pour trouver son chemin, une suite de clics dans un logiciel, une procédure administrative sont des algorithmes. Indiquer son chemin à un passant est une chose, acheminer ou prévoir le passage d'un flux massif de particules, de bactéries ou de clics web sont des procédures bien plus complexes. Pour les envisager, on fait appel à des outils algorithmiques existants, ce qu'on pourrait appeler des "familles algorithmiques" comme les graphes, les automates, les fractales, des modèles de la théorie des jeux, de l'émergence, etc. Il s'agit à la fois d'un dialecte pour décrire la réalité (le noeud, l'agent, la décision, le risque, le lien) et des outils mathématiques - des formules - et informatiques - des librairies de programmation - pour les traiter, analyser, modéliser.

Par exemple on utilise souvent les graphes pour traiter les données issues de réseaux sociaux : une personne est un noeud, une amitié est un lien, le nombre d'amis est le degré, une communauté est une certaine densité de liens entre des individus, etc. Avec le même formalisme, on peut tout aussi bien traiter des informations sur un réseau de gènes, sur des liens entre des pages web, etc. L'important est qu'une fois que les données sont définies selon ce formalisme, on a un ensemble d'algorithmes à notre disposition pour, par exemple, trouver le chemin le plus court d'un noeud à un autre, juger de la pertinence d'un noeud, mesurer la probabilité qu'un nouveau lien se forme à un endroit précis, définir des caractéristiques des données dans leur ensemble (centralité, diamètre, distribution de degré, etc). Cet ensemble d'outils a permis à une communauté de définir des éléments à un plus haut niveau d'abstraction comme, par exemple, la diffusion de l'innovation.

Dans ce contexte, un chercheur en algorithmique est souvent celui qui essaye de trouver de nouvelles combinaisons de ces outils informatiques, logiques et mathématiques pour améliorer des procédures ou en découvrir de nouvelles. Une bonne partie de la littérature dans ce domaine fait état des nombreux efforts consacrés pour réduire le "coût" d'une procédure. De la même manière, un séminaire de laboratoire peut typiquement avoir comme objet de retracer ces efforts dans le temps pour avoir une vue d'ensemble des tentatives passées, qui caractérise "l'expert" de ce problème. Ce coût est le plus souvent mesuré en temps, ie. le temps pris pour executer la procédure, et moins frequemment en espace, ie. l'espace mémoire nécessaire. Le temps est souvent formalisé avec la notation *Big O* (comparaison asymptotique en français) qui présente le nombre d'opérations nécessaire pour traiter n données. Par exemple, classer une liste de n noms par ordre alphabétique peut prendre selon la procédure choisie autant d'opérations que d'éléments dans la liste ($O(n)$), ou bien le carré de ce nombre ($O(n^2)$), etc.

À titre d'exemple, l'importance accrue donnée aux réseaux sociaux au milieu des années 2000 et aux données relationnelles qui leurs sont associées a rendu nécessaire la possibilité de calculer les communautés (groupe à fortes densité de liens) sur des graphes de très grande taille. Le Tableau 3 montre les principales étapes de cette recherche qui en moins de dix ans est passée d'un calcul impossible sur la plupart des corpus à des méthodes très efficaces sur des corpus massifs.

Une autre méthode pour chercher à réduire le cout d'un algorithme est d'en formuler une version "distribuée", c'est à dire réalisable par plusieurs machines en même temps. Au moins théoriquement, distribuer un calcul divise son coût en temps et en espace par le nombre de machines qui y contribue. Cependant nombre de procédures sont difficiles à formuler de la sorte, ce qui en fait un sujet de recherche à

Date	Algorithme	Complexité algorithmique	Nombre d'opérations
-	Force brute	$O(2^n)$	~ Impossible
2002	Girvan-Newman [40]	$O(n^3)$	1,000 milliards
2004	Fast Community [88]	$O(n^2)$	100 millions
2008	Louvain [14]	$O(n \log n)$	100,000

Tableau 3 – Quelques algorithmes pour résoudre la détection de communautés dans un graphe [15].

part entière dans les sciences informatiques. Comme nous le verrons dans le chapitre 2, le récent succès de certains algorithmes d'apprentissage est largement dû à la capacité de certains chercheurs, notamment KRIZHEVSKY, SUTSKEVER et HINTON [61], à implémenter ceux-ci sur des cartes graphiques destinées au jeu et au traitement d'image, qui permettent d'opérer un calcul en parallèle sur plus d'un millier de processeurs.

Ainsi, lorsque une procédure est éligible à une implémentation informatique, elle est susceptible de pouvoir être optimisée et profiter pleinement de la puissance de calcul toujours croissante. Ce qui rend éligible à cette optimisation une procédure, c'est le fait d'être suffisamment formalisée, précisée, dans des termes logiques et mathématiques qui permettent son exploitation. C'est notamment à ce titre que l'ont peut dire que la statistique est un fondement important de l'apprentissage, en ce qu'elle définit ce que formellement on entend par l'apprentissage depuis les données.

1.2.1.2 *Méthodes et disciplines de la statistique*

Les statistiques sont généralement considérées comme une science à part entière à partir du début du XX^{e} siècle, rassemblant ainsi un ensemble jugé cohérent de méthodes pour la mesure de l'incertitude, dans l'anticipation et l'interprétation des expériences et de l'observation. Avant que cette légitimité disciplinaire ne soit acquise, on observe plutôt comment des éléments de logique, communs à diverses sciences empiriques, ont émergés au croisement de plusieurs concepts mathématiques et des besoins de différentes sciences appliquées [128]. Plus simplement, les statistiques sont une technologie quantitative pour les sciences empiriques. Plus formellement, elle sont la logique de la mesure. Dans cette perspective, on pourrait considérer les statistiques comme une boîte à outils, un ensemble de ruses ou un recueil de techniques isolées, utiles à des activités scientifiques indépendantes.

En ce sens, la méthode des moindres carrés était dominante dans les mathématiques statistiques du XIX^e siècle. Elle est entendue comme un "calcul des observations", central à une statistique considérée avant tout comme "la combinaison d'observations". On attribue sa découverte à plusieurs personnalités, notamment Legendre (1752-1833), Gauss (1777-1855), Piazzi (1746-1826), qui ont formalisé et appliqué cette méthode dans un contexte principalement tourné vers l'astronomie - l'étude des astres - notamment pour ses applications à la géodésie, aidant à la navigation maritime. De la même manière, on retrouve plus tard l'émergence des méthodes de regression et de corrélation dans un autre contexte, celui de la bio-métrie, notamment pour l'étude de l'hérédité (Galton, Pearson). Les sciences sociales, politiques et juridiques adoptent ces méthodes notamment grâce au travaux de Adolf Quetelet (1796-1874) sur la représentation, et la représentativité, de "l'homme moyen" et des "causes persistentes" de son comportement. Ses études s'appuient notamment sur la loi binomiale et la loi de poisson.

Il y a donc une unité des méthodes statistiques qui reste reconnaissable, qu'il s'agisse d'une application en physique, sociologie, chimie, psychologie ou sciences politiques. Les quelques exemples cités ici nous permettent de voir, en amont de cette cohérence méthodologique amplement reconnue aujourd'hui, qu'intervient une histoire pourtant très disciplinaire de ces techniques. Les moments forts de cette "pré-histoire" des statistiques modernes sont donc composés à la fois de nouvelles techniques (moindres carrés, regression, probabilité inverse, etc), de nouvelles applications (Astronomie, Géodésie, Hérédité, Société), et d'objectifs généraux de ces méthodes (combiner des observations, mesurer l'incertitude, inférer de futurs évènements).

1.2.2 *Vers une nouvelle culture des données*

Il est difficile de pouvoir prétendre retracer le parcours exact des origines de l'apprentissage artificiel dans l'histoire complexe des statistiques qui ressemble plus à une géographie de contraintes et de découvertes qu'à une suite logique d'accumulation de savoirs. Il s'agirait alors de trouver parmi tous les grands courants de pensée qui constituent la construction de ce champ, un qui soit dès ces premières formulations, le plus proche d'une procédure d'apprentissage. En ce sens, les inférences bayésiennes et leur théorême éponyme sont une expression directe de la mise à jour d'un modèle, ou d'une opinion, à travers chaque élément de preuve qui se présente à lui (données). La simplicité de cette approche et son essence probabiliste et subjective vis-à-vis de ce qui peut être considéré comme vrai ou pas, en ont fait un objet controversé depuis ses origines au XVIII^e siècle et hante

encore aujourd'hui le débat entre fréquentistes et bayésiens dans de nombreuses communautés scientifiques. C'est donc en suivant le parcours de ce théorème pendant plus de deux siècles, en nous appuyant sur les travaux de McGrayne [76], que nous illustrerons la diffusion des idées et applications (§1.2.2.1) qui en construisant un terrain propice à l'usage des sciences informatiques, ont permis l'émergence d'une nouvelle culture statistique (§1.2.2.3) qui caractérise la science des données (§1.2.2.2).

1.2.2.1 *L'exemple des méthodes bayésiennes*

Les probabilités, entendues comme la mesure de l'incertitude, sont un des fondements de l'analyse bayésienne et apparaissent notamment dans le contexte de l'étude des jeux. Il s'agit pour bon nombre de ses précursseurs (Fermat, Pascal, Leibniz, Bernoulli, etc) d'étudier les mathématiques des permutations et des combinaisons afin de pouvoir énumérer les cas favorables d'une variété de jeux dont certaines propriétés sont connues. C'est notamment Simpson (1710-1761), Bayes (1702-1761) et Laplace (1749-1827), qui permettent de passer de l'énumération de cas favorables à l'inférence de cas futurs. Il s'agit alors de renverser la probabilité du comportement donné d'un système à la prédiction de son comportement futur. En partant de Pierre Simon Laplace et de ses recherches et applications sur les "probabilités inverses" (*inférence bayésienne*) - on peut suivre le parcours, sur plus de deux siècles, d'une nouvelle approche de la prédiction et observer ainsi comment elle jette les bases d'une nouvelle culture statistique.

Après des études de théologie, Pierre Simon Laplace travaille pour Jean d'Alembert principalement sur des questions d'astronomie, visant à étudier la stabilité de l'univers. Hébergé de manière assez précaire aux Invalides, et alors qu'il se voit refuser pour la sixième fois un poste à l'académie royale et envisage d'émigrer en Russie, Laplace découvre les travaux de Abraham de Moivre [81] qui avait inspiré Bayes quelques décénies auparavant. Dans un de ces premiers essais sur le sujet en 1774 [62], il considère alors l'approche probabiliste comme l'expression mathématiques de cette ignorance qui caractérise les humains, celle-là même qui fonde leur processus d'apprentissage, et offre ainsi la version contemporaine du théorème de Bayes.

Il présente à l'Académie des applications allant de scénarios de paris, à la forme de la planète terre, en passant par les mouvements de Jupiter et Saturne. Il étudie par la suite la disparité des naissances de garçons et filles et élimine certains déterminants admis à l'époque, comme le climat. Cette étude, pour laquelle il collecte des données en France, Angleterre, Russie, Egypte, Amérique Centrale, pendant plus de trente ans, lui sert aussi à prouver comment l'observation continuelle de données lui permet de se rapprocher à chaque instance

d'une certitude. De tels résultats légitiment ses méthodes pour l'établissement de politiques publiques et Condorcet lui demande, en ce sens, de participer à plusieurs réformes pour évaluer la population française, les procédures électorales, la fiabilité des témoins et jurés lors de procès.

Pour comprendre pourquoi Laplace développe ces méthodes alors que leur pionnier, Bayes, n'en avait pas fait d'application, il faut noter le contexte philosophico-religieux dans lequel ce dernier réalisa son travail : la recherche d'une cause ultime de tous les phénomènes, non sans référence religieuse à la présence d'un grand architecte. En opposition à cette approche, Laplace souligne l'intérêt pratique de ces méthodes en affirmant que "le vrai objet des sciences physiques n'est pas la recherche des causes premières mais la recherche de lois selon lesquelles les phénomènes sont produits" [63].

Mort en 1827, l'héritage de Laplace n'est pas bien considéré par ses successeurs qui lui reprochent son opportunisme politique auprès des différentes formes de gouvernement qui se succèdent de son vivant. Sur le plan scientifique, ils remettent en cause la paternité de ses découvertes, opinion largement entretenue pendant 150 ans, jusqu'au travaux de Stephen Stigler qui réintègre le mérite de Laplace [127]. Ces critiques ont souvent servi de base pour stigmatiser son travail sur les probabilités comme "une aberration de l'intellect et l'ignorance même inventé dans la Science" (John Stuart Mill, cité par Gigerenzer et Porter [38]) et son auteur comme "un des plus superficiels ayant obscurci l'histoire de la science" (Pearson, [97]). L'histoire qui s'en suit est principalement celle d'une condamnation théorique et généralisée des méthodes probabilistes de Laplace, qui influence la recherche et les publications académiques pendant près d'un siècle. En parallèle, l'utilité pratique des méthodes se réaffirme dans de nombreux contextes, notamment militaire, sans pour autant constituer un éceuil dans la doxa académique, largement occupée par le fréquentisme représenté notamment par Karl Pearson, Ronald Fisher et Jerzy Neyman.

Le bayésianisme survit donc en marge, au sein d'applications variées mais sans ambitions ou formulations théoriques. En France, le mathématicien Joseph Bertrand s'en sert pour étudier les stratégies de l'artillerie, reprises notamment pendant la première guerre mondiale par Jean Baptiste Estienne. On retrouve des références à ces méthodes dans le plaidoyer de Poincaré pour désavouer les preuves retenues contre Dreyfus in 1899. Au lendemain de la première guerre mondiale, plusieurs initiatives font usage de ces méthodes aux États-Unis. Par exemple, Edward Molina utilise ses méthodes pour gérer le routage des communications téléphoniques, mais ses travaux ne sont publiés qu'en interne et ne sont rendus publics que tardivement par brevet ([82]). Aussi, Isaac Rubinow collecte des données sur les acci-

dents du travail pour étudier leurs causes et leurs probabilités futures, jetant ainsi les bases de l'assurance sociale. Les méthodes bayesiennes ont eu un rôle important dans le déchiffrement par l'armée anglaise des communications de l'armée nazi lors de la seconde guerre mondiale. Alan Turing et les autres mathématiciens impliqués dans ce projet explorent alors les premières interactions entre méthodes statistiques et pré-informatiques notamment en utilisant des méthodes bayésiennes pour réduire le nombre de possibilités devant être explorées par ce qui est considéré comme un des premier ordinateur (le colossus). Cependant le fruit de ces recherches en temps de guerre étant couvert par le secret, la plupart de ces découvertes et les traces de l'usage des méthodes bayésiennes sont détruites, laissant le succès de ces méthodes dans la même situation controversée et isolée qu'avant-guerre.

La situation marginale du bayésianisme se reproduit donc encore pendant plusieurs décénnies, à tel point que le premier article pratique expliquant aux scientifiques comment utiliser ces méthodes date de 1963, note McGrayne [76]. On assiste dans cette période, comme précedemment, à des usages fructueux mais isolés ou secrets qui maintiennent l'intérêt d'une communauté, mais en marge de l'académie. Parmi les quelques succès de cette époque, on trouve de nouvelles applications au calcul des assurances (Arthur Bailey), les premiers arguments statistiques faisant le lien entre cigarettes et cancer du poumon (Hill et Doll), des méthodes décisionnelles pour le management (Robert Osher Schlaifer et Howard Raiffa), et bien d'autres encore.

L'avantage principal qui apparaît dans ces exemples est la capacité des méthodes bayésiennes à pouvoir émettre des hypothèses même lorsque les données dont on dispose comportent de nombreuses inconnues. John Tukey (1915-2000) va dans ce sens en critiquant l'austérité du fréquentiste Fisher en affirmant que "il est préférable d'avoir une réponse approximative a une bonne question, qu'une réponse exacte à une mauvaise question". C'est cette capacité probabiliste et adaptable à de nombreux contexte qui, selon McGrayne [76], consacre les méthodes bayésiennes lors de l'avènement de l'informatique, notamment l'informatique d'entreprise, et la possibilité accrue d'expertise basée sur des données. Cette rencontre entre statistique et informatique donne ces lettres de noblesses au bayésianisme. Cette méthode est privilégiée car elle tire partie de la puissance de calcul et des données peu structurées qui deviennent de plus en plus accessibles. Cependant, cet intérêt commun entre techniques statistiques et informatique, s'il a servi le bayésianisme, n'a été, en grande partie, formulé ainsi que par la naissante science des données.

1.2.2.2 *Data Science*

L'idée de science des données est attribuée à Peter Naur qu'il définit comme "la science qui traite des données, une fois obtenues, alors que la relation des données avec ce qu'elles représentent est déléguée à d'autres domaines scientifiques" [16] [85]. Il s'agit donc d'un substitut au mot informatique, l'auteur étant connu par ailleurs pour ne pas vouloir réduire à la pratique formelle de la programmation ou à une branche des mathématiques. En ce sens, il décrit son champ disciplinaire comme traitant de l'essence de la rencontre entre les problèmes, les outils, et les gens :

> "On ne peut pas, en tant que personne, penser un problème sans au même moment inclure un type d'outil. De plus, quand l'outil change, le problème n'est plus le même. Inversement, notre opinion sur ce qu'est un outil adéquat dépend de notre compréhension du problème. Dans tous les cas, le problème et l'outil ne sont rien s'ils ne sont pas reconnus par une personne - c'est là que les gens prennent leur importance." [17]
>
> NAUR [84]

Ainsi, le traitement de données qu'implique la résolution de problèmes pour les gens, semble être une première définition possible de la science des données. Il s'agit donc d'ordonner, du latin ordinare, "mettre en ordre, arranger", à dessein. Alors que l'idée que cela implique des gens semble acquise à l'usage ancien du mot ordinateur ("celui qui est chargé de régler les affaires publiques" [18]), la discipline informatique s'en empare d'abord peu et se concentre sur une représentation agnostique des données et des situations à optimiser, comme la vitesse d'exécution, la quantité de mémoire utilisée, pour, par exemple, classer une liste, chercher un élément dans celle-ci, etc.

La science des données est associée dans un deuxième temps aux statistiques, notamment par WU [147], qui propose de confondre les deux et de nommer son praticien le *data scientist*. Cette proposition se fait non sans rappeler les premiers usages du mot statistique, par

16. "Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences."

17. "We cannot, as people, think of a problem without at the same time implying some kind of tool. Stronger yet, when the tool changes the problem is not the same any more. On the other hand, our opinion about what is a proper, or desirable, tool surely depends on our understanding of the problem. In any case the problem and tool are nothing if they are not recognized as such by a person–that is where the people come in."

18. *in* XVI^{e} siècle-début XVII^{e} siècle. (Pasquier, Lettres, II, 5 ds Hug.)

exemple dans un manuel de 1770 où les statistiques sont dites rassembler les techniques qui "nous apprennent l'arrangement politique de tout les états modernes du monde connu" [11]. Plus de deux siècles après, Ronald Fisher rappelle que "le sens original du mot statistique suggère qu'il s'agissait de l'étude des populations d'humains vivant en union politique" [19] [36]. Dans le même sens, quelques années plus tôt, Karl Pearson prête de nombreuses missions sociétales aux statistiques dans *La grammaire de la science*, surtout vis-à-vis de la loi juridique que le calcul statistique pourrait rapprocher des lois scientifiques [7, 96].

Comme pour l'analogie avec l'informatique, on trouve l'idée qu'il s'agit de problèmes impliquant des gens et leurs sociétés ou communautés. La définition des statistiques qui accompagne celle de la science des données dépasse largement celle des statistiques descriptives. Il s'agit d'une approche où la statistique sert des décisions sur la base des connaissances extraites des données, et traite en amont la manière de les extraire.

On attribue souvent à John Tukey [135, 136] la formulation d'une vision suffisamment empirique sur les statistiques qui permet l'association entre les sciences informatiques et la théorie de l'information. Cet ensemble d'outils, en partie allégé du poids de leur tradition théorique, peut être envisagé avec les contraintes de domaines d'applications nouveaux. En ce sens, en 1977, l'Institut International de Statistique (ISI) fonde l'Association Internationale de Statistiques Computationnelles avec pour but de "lier la méthodologie statistique traditionnelle, les technologies informatique modernes, et les connaissances d'experts afin de transformer les données en information et connaissance" [20]. Cette initiative institutionnelle est suivie par de nombreuses autres au cours des décennies qui suivront. Le Tableau 4 reprend quelques-unes de ces étapes.

Ces éléments d'une histoire linéaire et récente de la science des données ne doivent pas nous faire sous-estimer l'importance des autres termes ayant chercher à subsumer l'ensemble des intérêts en mouvement dans l'analyse des données. En ce sens, le *knowledge discovery*, le *data mining*, *data analysis* ont été aussi des tentatives de rassembler un ensemble de pratiques et de mutations de la recherche et des développements scientifiques et industriels. D'autres niches témoignent de l'attrait pour des approches interdisciplinaires de l'analyse des données. Par exemple, les systèmes complexes, les sciences sociales computationnelles [64], la socio-physique, l'analyse de réseaux sociaux,

19. "The original meaning of the word *statistics* suggests that is was the study of populations of human beings living in political union."

20. "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

1989	Gregory Piatesky-Shapiro organise le premier atelier de *Knowledge Discovery in Databases* (KDD), qui reste aujourd'hui un des rendez-vous important du mouvement "Big data".
1994	*Business Week* publie un dossier spécial sur le *Database Marketing* où il décrit des "entreprises qui collectent des montagnes de données et d'informations à propos de vous, pour savoir à quel point vous êtes prêt à acheter un produit" [21].
1995	KDD est intégré à l'*Association of Computing Machinery* (KDD).
1996	la Fédération Internationale des Sociétés de Classification (IFCS) se réunit au Japon et titre sa rencontre "Data Science, classification and related methods".
1997	Le journal *Data Mining and Knowledge Discovery* est lancé.
2001	William Cleavland des Bell labs publie un rapport intitulé : "Data Science : Un plan d'action pour étendre les domaines d'applications de la statistique" [24].
2002	Lancement du *Data Science Journal*.
2003	Lancement du *Journal of Data Science*.
2005	le Comité National Scientifique des EUA publie un rapport mettant en avant le rôle de la science des données dans l'éducation [122].
2007	L'université de Fudan (Shangai, Chine) lance le *Research Center for Dataology and Data Science*, en déclarant qu'il s'agit d'une nouvelle science à part entière [151].
2012-2015	Des initiatives similaires sont mises en place aux EUA, notamment à Columbia (2012), NYU (2014), MIT et l'Université du Michigan (2015). En France, de tel cursus voient le jour dans les écoles Polytechnique et Centrale en 2014.

Tableau 4 – Quelques étapes importantes de l'émergence institutionnelle de la science des données.

l'analyse quantitative, etc, sont autant de termes avec d'une part un ensemble de méthodes privilégiées, et, d'autre part, une volonté de subsumer d'autres techniques d'analyse. À ce dernier jeu, il semble que le terme de "Data Science" ait remporté le droit de labéliser les autres, en témoigne comment ces derniers se redéfinissent dans les termes des sciences des données. Une piste pour expliquer ce succès serait que parmi toutes ses dénominations, *Data Science* est la plus neutre, n'impliquant aucun contexte disciplinaire particulier, laissant donc ainsi la place à des appropriations plus diverses.

L'apprentissage automatique fait justement partie des disciplines dont l'identité académique et industrielle devient fortement liée à celle de la science des données. Ce lien est assez naturel du fait que le *machine learning* est dépendant de données pour progresser, à tel point que certains chercheurs considèrent que la constitution de base de données d'apprentissage est de très loin plus importante à l'évolution de cette discipline que l'amélioration des algorithmes ou l'extraction de variables pertinentes [17]. Mais il semble que ce qu'incarne le plus l'apprentissage artificiel dans le contexte de la science des données c'est le fait de proposer une nouvelle statistique qui abstrait ses méthodes dans le sens d'une efficacité mesurable, optimisée, qui délaisse l'observation et l'explication que celles-ci peuvent offrir.

1.2.2.3 *Une nouvelle culture de modélisation statistique*

L'usage intensif des outils informatiques et une nouvelle appropriation de la statistique permettant au *machine learning* de se construire une identité propre que Léo Breiman a tenté d'analyser sous le prisme d'une culture émergente de la modélisation. Ainsi, dans son article *Statistical Modeling : The two cultures* [20], il reprend les principaux points de la dissension au sein de la communauté statistique au début des années 2000. Pour lui l'apprentissage artificiel est identifié comme une culture émergente de la pratique statistique, qu'il qualifie "d'algorithmique" en opposition à celle "des données", plus classique. Ce qui différencie principalement ces deux approches c'est la manière dont elles considèrent que les données témoignent du phénomène étudié.

L'approche traditionnelle considère son objet d'étude comme un processus dont les paramêtres sont estimés au regard du modèle généré par la procédure statistique. C'est à dire qu'elle considère le modèle statistique comme étant celui à l'origine du phénomène "naturel". Dit plus simplement, on accorde un certain degré de "vérité" au modèle statistique et à ce titre on peut donc observer et discuter du modèle et de ses paramètres comme s'il s'agissait du phénomène lui-même.

L'approche algorithmique considère le phénomène étudié comme inconnu et le modèle généré ne cherche qu'à augmenter la probabilité qu'une classification soit pertinente ou qu'une prédiction soit juste. Dit autrement, l'approche algorithmique ne considère pas que les données sont générées par son modèle. Le modèle est un moyen distinct du phénomène "naturel" étudié et ne prétend représenter aucune "vérité" à son égard. Il est établi pour augmenter la pertinence d'une action sur des comportements futurs de ce phénomène.

Afin d'illustrer cette dichotomie et l'étendue de ses implications, nous prendrons l'exemple d'un domaine d'application classique de l'histoire des statistiques, l'assurance automobile. Une compagnie d'assurance automobile vise notamment à garantir qu'en cas d'incident les personnes concernées soient indemnisées des dommages reçus. La conduite automobile est associée à une notion de "bonne conduite" qui fait varier le prix de l'assurance en fonction de ce que l'on peut attendre, ou prédire, du comportement d'une personne en fonction de critères le concernant (âge, années de conduite, etc) ou de son comportement (nombres d'accidents fautifs, de contraventions, etc).

Une approche "traditionnelle", autant pour la statistique que pour l'assurance, est d'utiliser ces différents critères afin de constituer, ou apprendre, un modèle qui définira le prix de l'assurance à partir de la prédiction du comportement de l'assuré. Typiquement, ce modèle est issue d'une analyse statistique proche de celle décrite précédemment (Tableau 2) dont le résultat pourrait grossièrement ressembler à :

$$
\begin{aligned}
&3 \times \text{nombre d'accidents fautifs} + \\
&6 \times \text{nombre de contravention pour conduite en état d'ivresse} + \\
&10 \times \text{age} + \\
&5 \times \text{nombre de chevaux du modèles de la voiture} \\
&= \text{Prix de l'assurance}
\end{aligned}
\tag{1}
$$

Il s'agit donc des critères retenus pour chaque assuré, auquel l'algorithme d'apprentissage attribue des poids en fonction de leur importance dans le modèle. L'implémentation de ce modèle dans les procédures de l'assurance et le calcul du prix pour chaque client fait l'objet de nombreuses négociations tant en amont qu'en aval de l'analyse statistique. En effet, la prise en compte de certains critères est interdite (par exemple, la classification ethnique), d'autres font l'objet de négotiations avec des associations de consommateurs ou des initiatives publiques, par exemple l'augmentation des pénalités pour la conduite en état d'ébriété. Ce modèle a donc une responsabilité juridique, sociale, politique auquel peut s'ajouter une responsabilité

économique ou financière si l'on considère que les investisseurs de l'entreprise d'assurance souhaitent que ce modèle reste performant vis-à-vis de ceux de ses concurrents.

Ces responsabilités diverses, indirectes au problème que le modèle tente de cerner, ont aussi des repercussions sur la répartition du travail que demande ce modèle. Vu l'importance des paramètres du modèle vis-à-vis de la loi ou des investisseurs, on va souvent décider quels paramètres doivent être mis en valeur en amont de la procédure statistique. Une hiérarchie décisionnaire contraint les statisticiens de l'entreprise d'assurance à aller dans un sens plutot qu'un autre, favoriser une représentation des données et des risques. Si ce modèle avait pour objectif d'analyser l'insécurité routière, on pourrait juger de la place de l'inexpérience du conducteur ou de l'usage de l'alcool au volant dans les accidents, etc. Ce modèle et ses paramètres représentent le phénomène qu'il modélise et rentre ainsi tout à fait dans la définition de la modélisation statistique "traditionnelle" définie par Breiman.

À l'opposée de cette approche, et afin d'illustrer ce que Breiman appelle "l'approche algorithmique de la modélisation statistique", on prendra l'exemple d'une entreprise française proposant une nouvelle forme d'assurance automobile. Cette entreprise, *YouDrive*, invite ses clients à installer dans leurs voitures un boitier gratuit muni d'un accéléromètre et d'un GPS. Ce boitier permet à l'entreprise de juger de la conduite réelle de ses clients et ainsi d'éviter les pénalisations sur des malus (critères) définis *a priori*. Parmi les malus que *Youdrive* permet de contourner, l'âge est le principal mentionné dans leurs diverses publicités. En effet, si l'âge est une variable permettant de prédire une conduite novice ou impétueuse, c'est parce que certains types de conduite à risque sont plus probables chez les jeunes que chez les conducteurs expérimentés. Mais si mesurer directement la qualité d'une conduite devient possible, la déduction lié au critère de l'âge devient moins pertinente.

La boite noire plaçée dans la voiture des clients de *YouDrive* doit probablement produire une longue séquence de données. Au lieu de n'avoir que quelques variables pour chaque client, *YouDrive* dispose de toutes les données de mouvement qui captent beaucoup d'informations de premières main comme l'accélération, le freinage, comment les virages sont pris, l'allure, etc. Tout cela peut être corrélé avec le GPS qui indique la position géographique et qui permet de comparer un client à un autre pour observer, par exemple, si une réaction brusque est prévisible à un certain endroit. Il est impossible de prétendre savoir ce que fait vraiment la méthode d'apprentissage appliquée à ces données mais on peut néanmoins explorer les possiblités offertes à l'entreprise du fait de la masse de données dont elle dispose. En effet, on peut imaginer qu'un algorithme d'apprentissage

utilise l'ensemble des données issues de l'accéléromètre et du GPS pour déterminer le prix de l'assurance sans avoir de paramètres, ou de critères clairs et explicites, car le modèle produit est extrêmement complexe. Cette complexité peut être représentée par un modèle qui aurait pour paramètre, par exemple, chaque seconde de conduite, rendant impossible l'émergence de variable lisible.

Ici le modèle perd son pouvoir explicatif. La représentation de la réalité, pourtant beaucoup plus fidèle et individualisée que pour un modèle classique, n'est plus orientée par quelques variables claires mais par un nombre important de variables choisies exclusivement comme déterminantes à l'apprentissage et pour l'optimisation de la prédiction. Il s'agit ici, dans les termes de Breiman, d'une approche algorithmique de la modélisation, c'est à dire que l'accent est mis davantage sur la performance prédictive de la méthode que sur la représentation intelligible de la réalité qu'elle modélise. Ce type de modèle ne peut pas être négocié dans les mêmes termes que le précédent. On peut difficilement y opposer une variable, comme la conduite en état d'ébriété, puisque le modèle n'en a pas vraiment (on dit qu'il est non-paramétrique). Ainsi les responsabilités sociales, politiques ou juridiques observées dans l'approche traditionnelle ne peuvent lui être réclamées directement puisque leurs critères de responsabilité ne peuvent être observés directement. Enfin, ces critères étant inexistant dans un modèle complexe, l'entièreté du modèle repose, dans la hierarchie de l'entreprise, beaucoup plus sur la personne qui réalise la procédure d'apprentissage que sur un ensemble de décisions prises en amont pour définir le modèle.

Pour reprendre les principaux traits de la dichotomie de Breiman dont on a voulu donner un exemple ici, on peut dire que : l'approche "traditionnelle" peut se baser sur peu de données et peu de critères, produire un modèle intelligible, interprétable, soutenir une responsabilité sociale, politique et donc des décisions, choix, qui sont fait en amont de sa réalisation. L'approche "algorithmique" nécessite beaucoup de données et de critères pour chaque observation. Le modèle qu'elle produit est peu ou pas interprétable et ne peut donc pas être tenu pour responsable de ses critères qui ne sont pas apparents. En contrepartie, il est généralement beaucoup plus performant, individualisé et moins dépendant de moyennes et d'indicateurs généraux.

La typologie de Breiman vise explicitement à rendre compte de l'émergence, au début des années 2000, d'une communauté de chercheurs et d'ingénieurs qui se concentre sur une approche plus pragmatique et performante des méthodes d'apprentissage et de la modélisation statistique. En appréhendant cette communauté par cette distinction, Breiman nous offre un premier moyen de représentation du *machine learning* que le chapitre suivant tente de compléter en observant et analysant plusieurs des algorithmes qui le composent.

RÉSUMÉ DU CHAPITRE 1

Dans ce chapitre, nous avons déroulé plusieurs exemples de procédures d'apprentissage artificiel afin de développer une première intuition de comment ce domaine permet de classifier des données ou appuyer des processus de découverte et de prédiction à partir de celles-ci.

Ces premiers exemples nous ont permis de définir l'ambition de l'apprentissage par la négative, en l'opposant à la pratique de la programmation et du *design* dont l'architecte est censé extraire lui-même les motifs des données et les assembler logiquement pour implémenter un processus de décision. De manière plus constructive, l'apprentissage est apparu avant tout comme l'imitation du processus éponyme chez l'humain et l'animal, et l'hypothèse principale que soutient cette recherche est celle qu'un processus unique peut résoudre bon nombre de problèmes de manière générique.

Si l'apprentissage artificiel semble obtenir une cohérence d'ensemble et ses lettres de noblesse lors des dernières décennies, on peut faire remonter son héritage aux premiers efforts de la statistique et des probabilités depuis le XVII^{e} siècle. Ses caractéristiques plus récentes ont à voir avec l'informatique, l'usage de la puissance de calcul et le rapprochement naturel qui se construit avec ladite "science des données", qui contribuent à formuler une nouvelle culture de modélisation des données.

2

TYPOLOGIE DES PROCÉDURES D'APPRENTISSAGE ARTIFICIEL

SOMMAIRE

Dans le chapitre 1 nous avons pu voir que l'apprentissage artificiel incarne une nouvelle culture de modélisation statistique en empruntant des références à plusieurs traditions scientifiques desquelles elle se différencie par la proposition singulière : comment résoudre un problème ou extraire de la connaissance à partir de données. Ce chapitre adopte un point de vue plus internaliste en cherchant à montrer comment cette ambition est partagée par différents courants de pensée aboutissant à des techniques, méthodes et algorithmes différents et ayant leur propre trajectoire.

Pour ce faire, chaque section de ce chapitre accompagne le lecteur dans une description détaillée des principaux algorithmes d'apprentissage en identifiant à chaque fois leurs principes d'inférence. Ce chapitre ne prétend donc pas présenter un état de l'art exhaustif des algorithmes d'apprentissage artificiel, ni à en décrire le fonctionnement fin, mais il voudrait plutôt fournir à un lecteur néophyte une intuition de chacune de ces *épistémès* et montrer comment elles proposent chacune une solution propre à des problèmes similaires tout en s'appuyant sur des métaphores différentes. Ainsi, chaque section accompagne la description des algorithmes concernés d'éléments historiques sur sa formulation et son évolution et parcourt plusieurs problématiques transversales comme la place faite à l'intelligibilité et l'interprétabilité des modèles statistiques produits, leur propension à l'erreur et au sur-apprentissage, leur capacité à être implémenté de manière distribuée, etc.

Nous traitons successivement des arbres de décision et forêts aléatoires (§2.1), des réseaux bayésiens (§2.2), des algorithmes génétiques (§2.3), des machines à vecteurs de support (§2.4) et des réseaux de neurones artificiels (§2.5). Ainsi, nous verrons que chacune de ces communautés propose un *style de raisonnement* caractéristique d'une certaine forme d'apprentissage artificiel. Ces styles font l'objet d'une typologie et d'une analyse commune dans la dernière section de ce chapitre (§2.6).

Le choix de traiter de cette liste d'algorithmes est fortement ancré dans l'exploration des multiples matériaux pédagogiques et de recherche rencontrés pendant cette thèse. Cependant, pour y donner plus de poids qu'une simple intuition de leur représentativité, on peut se tourner vers les récents travaux de Pedro Domingos [33]. En effet, dans *The Master Algorithm*, cet auteur reconnu dans la communauté de l'apprentissage entreprend de vulgariser son expertise en présentant les familles d'algorithmes qu'il juge représentative des efforts entrepris par le *machine learning*. Cette typologie que nous détaillons davantage dans la dernière section (§2.6) s'appuie sur la même sélection d'algorithme que ce chapitre.

2.1 ARBRES DE DÉCISION ET FORÊTS ALÉATOIRES

Un arbre de décision est une représentation formelle des critères menant à une prise de décision et, à ce titre, tire ses origines des systèmes experts, en offrant une façon de représenter une séquence de décisions qu'un expert aurait pu prendre face à une question donnée. C'est donc une méthode ancienne pour représenter les chances qu'a un évènement de se produire, le coût des ressources engagées, l'utilité de celles-ci, etc. C'est aussi une manière assez naturelle de représenter un algorithme et son comportement en fonction de variables et de contextes différents. En ce sens, un arbre de décision peut toujours être traduit comme une succession de boucles conditionnelles du type $\texttt{if} - \texttt{then} - \texttt{else}$ qui manipulent un certain nombre de symboles représentant les divers états possibles des variables du système considéré. On peut, par exemple, exprimer l'expertise du naturaliste face à une fleur, dont il peut mesurer la taille des pétales, sous la forme de l'arbre de décision suivant dont les feuilles aboutissent aux différentes espèces possibles.

L'évidence de cette représentation, tant pour un expert que pour un algorithme, rend difficile de faire l'histoire de cette technique et d'isoler ses origines en tant que formalisme générique modélisant la prise de décision. Néanmoins, parmi les algorithmes d'apprentissage, les arbres de décision peuvent être identifiés à la famille de ceux qui "divisent pour régner" (*divide and conquer*). Ainsi, de manière récursive,

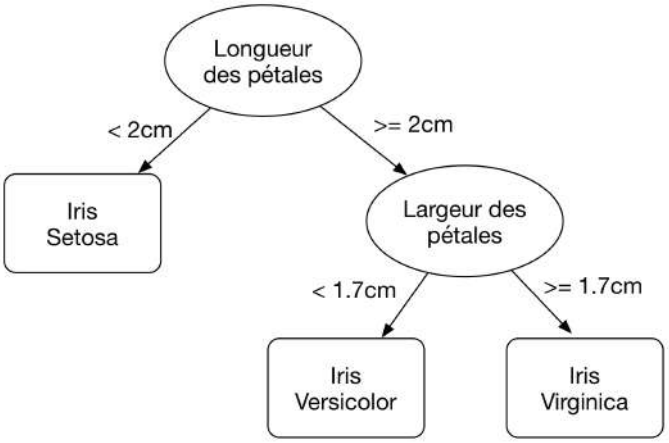


FIGURE 6 – Exemple minimaliste d'arbre de décision sur la base de donnée Iris [1]

la procédure d'un arbre de décision divise un problème en plusieurs sous-problèmes qui sont similaires au premier, mais de moindre taille. La réunion de ces bisections successives assure la résolution du problème. Il s'agit donc d'identifier les sous-espaces de l'espace de variables initial partageant la même solution.

Si l'usage d'un arbre de décision est assez intuitif pour un expert visant à transmettre sa connaissance d'un problème, dans le cas d'une procédure d'apprentissage, il faut trouver une méthode systématique qui détermine les questions pertinentes à poser, dans le meilleur ordre possible, de manière à optimiser la taille de l'arbre, et ainsi le temps de calcul associé à son exécution. Le même enjeu se pose dans le jeu des 20 questions où un joueur doit deviner à quelle personne un autre joueur pense en posant une série de questions fermées (en toute généralité, les arbres de décision ne sont a priori pas limités à des choix dichotomiques). En admettant que les réponses à toutes les questions sont positives, une partie pourrait prendre la forme suivante :

- Est-ce un homme ? Oui
- Est-il mort ? Oui
- Était-il un scientifique ? Oui
- Était-il engagé politiquement ? Oui
- Était-il démocrate ? Oui
- A-t-il contribué à la formulation des forêts aléatoires ? Oui
- **Décision** : Léo Breiman

On peut se représenter la stratégie du joueur comme suit. Partant de sa base de données mentales composée de l'ensemble des choses qu'il connait et de leurs caractéristiques, ou variables (métier, engagement, taille, sexe, etc.), la série de questions à laquelle il soumet

l'autre joueur vise à séparer cet espace le plus efficacement possible : entre morts et vivants, homme ou femme, etc... Le but étant de poser le nombre le plus petit possible de questions. Chaque question fait référence à une variable qui devient alors, à cette position dans l'arbre, le meilleur classifieur.

Une des premières méthodes pour qualifier l'efficacité de chaque question a été formulée par BREIMAN et al. [19]. Ces auteurs utilisent le coefficient de Gini, nommé en référence au sociologue Corrado Gini qui publie en 1912 [39] cet indicateur pour rendre compte des inégalités de revenus d'une population. Si cette mesure de la dispersion des données permet effectivement de prédire la pertinence d'une question, elle est rapidement supplantée par un emprunt à une autre tradition scientifique, la théorie de l'information. En effet, en 1986, Quilan publie une nouvelle méthode, ID3 [101], qui repose sur le gain d'information, c'est à dire la réduction d'entropie causée par la division des données par la variable considérée. Par réduction d'entropie on entend la réduction d'aléatoire dans les données, ou, inversement, l'augmentation de leur prédictibilité. L'usage de l'entropie, comme le fait ID3 et son successeur C4.5 [102], est la méthode la plus envisagée aujourd'hui pour construire un arbre de décision. Cet emprunt à la théorie de l'information illustre bien comment l'apprentissage artificiel a pu s'enrichir d'héritages scientifiques assez divers, de la sociologie aux mathématiques, en définissant ses objectifs propres, ici, l'efficacité d'une procédure inductive d'apprentissage.

La représentation d'un arbre de décision appris est donc identique à un système expert. C'est pour cela que cette méthode est particulièrement utilisée pour obtenir un modèle transparent, directement interprétable, c'est à dire dont la procédure de décision déclare explicitement quelles sont les variables qui guident sa prédiciton. Ainsi, certains domaines d'application en font un usage privilégié notamment lorsque les décisions appellent une justification explicite. MITCHELL [80], dans son livre d'introduction au *machine learning*, affirme ainsi que le succès de ces méthodes pour les diagnostics médicaux et l'attribution de crédits bancaires s'expliquent par cette propriété.

Cet lisibilité du cheminement de la décision à travers les différentes branches explique pourquoi l'arbre de décision est souvent invoqué comme méthode pour analyser une décision, en rendre compte, la comparer, analyser ses coûts et bénéfices, etc. Il est plus facile pour une personne, par exemple un médecin, ou une institution, par exemple un hopital, d'accepter d'en prendre la responsabilité, peu importe qu'il s'agisse d'un modèle appris ou d'un système expert. Pour des domaines comme le diagnostic médical ou l'attribution de crédit bancaire, pouvoir justifier ses décisions est souvent une obligation légale, ce qui peut expliquer le succès des méthodes lisibles et symboliques comme les arbres de décision dans ces domaines. En ce sens, la simi-

larité entre arbre de décision appris et la représentation en système expert illustre très bien l'approche traditionnelle de la modélisation statistique [2] dont le modèle est pensé comme un moyen de discuter de la réalité qu'il représente. Néanmoins, cette simplicité opérationnelle et interprétative comportent plusieurs défauts qui pénalisent la performance de la procédure d'apprentissage, invitant ainsi la recherche académique sur ces procédures à explorer de nouveaux compromis entre performance et interprétabilité.

Un des défauts des arbres de décision est qu'ils sont sensibles au sur-apprentissage, c'est à dire qu'il peuvent facilement épouser de manière excessive les données d'apprentissage et perdre ainsi en capacité de prédiction lorsque de nouvelles données lui sont présentées. Moins formellement, si les arbres de décision parviennent à dégager des règles générales, ils s'encombrent de détails inutiles. On dit alors qu'ils *généralisent* mal. Une technique pour aller à l'encontre de ce défaut est l'élagage, qui consiste à considérer chaque noeud de l'arbre comme une potentielle feuille (une extrémité) qui assignerait la classe la plus probable de la structure élaguée [78].

Un autre défaut, propre à toute démarche d'apprentissage, est que si l'on considère un jeu de données, il existe une multitude d'arbres de décision permettant de les décrire dans une démarche prédictive. Selon les algorithmes, certains types d'arbres seront privilégiés. Par exemple, ID3 aura tendance à privilégier les arbres courts, moins complexes. Ce biais d'ID3 est souvent justifié par le principe philosophique et scientifique du *rasoir d'occam* selon lequel les hypothèses les plus courtes sont les meilleures dans la mesure où elles rendent compte tout aussi bien des données qu'un modèle plus complexe. Cependant, si la présence de nombreuses hypothèses possibles inquiètent sur le fait qu'une hypothèse ne soit pas forcément la meilleure, cela permet aussi d'envisager de tirer partie de l'ensemble de celles-ci. En ce sens, les techniques de *bagging* visent à entrainer différents modèles sur des variations légères du corpus de données initial, et de les combiner en faisant une moyenne de celles-ci. Appliquer aux arbres de décision, Leo Breiman a baptisé cette technique "forêts alétoires" [18] qui, en plus d'utiliser des variations du corpus de données initiaux, utilise des ensembles réduits des variables qui diffèrent pour chaque arbre. Si cette technique offre des performances accrues en terme de prédiction et de classification, elle éloigne aussi le modèle des systèmes experts et de ses vertus d'interprétabilité.

Leo Breiman développa cette méthode lorsqu'il quitta l'université pour travailler comme consultant pour des projets industriels et de politiques publiques. Il travailla notamment pour l'agence de protection de l'environnement américaine sur différents modèles prédictifs sur la couche d'ozone et la toxicicité de certains composés chimiques,

2. cf. §1.2.2.3

et pour d'autres clients sur la reconnaissance d'images et de voix. Ces missions exigeaient d'obtenir le plus de précision prédictives possible et c'est dans le cadre de ces contraintes qu'il développa les forêts aléatoires.

Alors qu'il publie son article sur les forêts aléatoires, Breiman écrit - la même année - son article sur les "cultures de modélisation" [20] où il prend justement comme exemple cette nouvelle technique pour illustrer la culture algorithmique qui émerge au sein des statistiques. Il montre comment augmenter la complexité du modèle, comme le font les forêts alétoires, implique une moindre interprétabilité de celui-ci. La multiplicité des modèles générés par les différents jeux de données moins apte à expliquer ou comprendre pourquoi telle ou telle prédiction a été produite. Cet état de fait est symptomatique de ce qu'il décrit comme un compromis entre précision et interprétabilité, un complément au principe de rasoir d'Ockham, qui permet de s'adapter à différents objectifs et contraintes.

2.2 RÉSEAUX BAYÉSIENS

Dans le chapitre précédent (§1.2.2.1) nous avons déjà eu l'occasion de retracer un historique rapide du développement et des usages des méthodes bayésiennes en statistiques, depuis Pierre Simon Laplace, en 1774 [62], jusqu'à leurs appropriations contemporaines par l'informatique et la science des données. Les approches bayésiennes en apprentissage artificiel sont une extension du théorème de Bayes à des contextes informatisés : des corpus massifs de données et un nombre élevé de variables. Il s'agit davantage d'améliorations, d'implémentations, et de nouveaux usages du théroème de Bayes, que de nouvelles hypothèses mathématiques. Le théorème lui-même permet d'inverser une probabilité. Son utilisation par les méthodes d'apprentissage permet l'établissement d'un modèle probabiliste pour la prédiction ou la classification.

Imaginons que nous souhaitions construire un modèle pour estimer la probabilité qu'un article de presse parle du réchauffement climatique. Pour ce faire, nous constituons un corpus d'articles de presse pris au hasard dont nous savons lesquels parlent du climat ou pas. De plus, pour chacun des articles, nous construisons deux variables binaires selon qu'ils contiennent ou pas les mots "taxe" et "carbone". L'idée de cette procédure d'apprentissage est de pouvoir notamment saisir la forte probabilité qu'un article parle de rechauffement climatique s'il contient les mots "taxe" et "carbone", sans attribuer la même probabilité à leur occurence isolée. La zone verte de la Figure 7 rassemble les éléments dont nous disposons sur ces données, à savoir :

- la probabilité qu'un article traite du réchauffement climatique, $\mathrm{P}(climat)$,
- celle qu'un article contiennent le mot "taxe", $\mathrm{P}(taxe)$, et
- celle qu'un article contienne le mot "carbone", $\mathrm{P}(carbone)$.

En parcourant le corpus, on peut aussi énumérer le nombre d'articles qui contiennent ces mots parmi ceux qui traitent du réchauffement climatique. Ainsi on peut connaitre les effets (zone rouge) de l'appartenance à la classe que nous essayons de modéliser, c'est à dire :

- la probabilité qu'un article mentionne le mot "carbone" sachant qu'il parle du climat, $\mathrm{P}(\text{“}carbone\text{”}|climat)$, et
- celle qu'un article mentionne le mot "taxe" sachant qu'il parle de climat, $\mathrm{P}(\text{“}taxe\text{”}|climat)$.

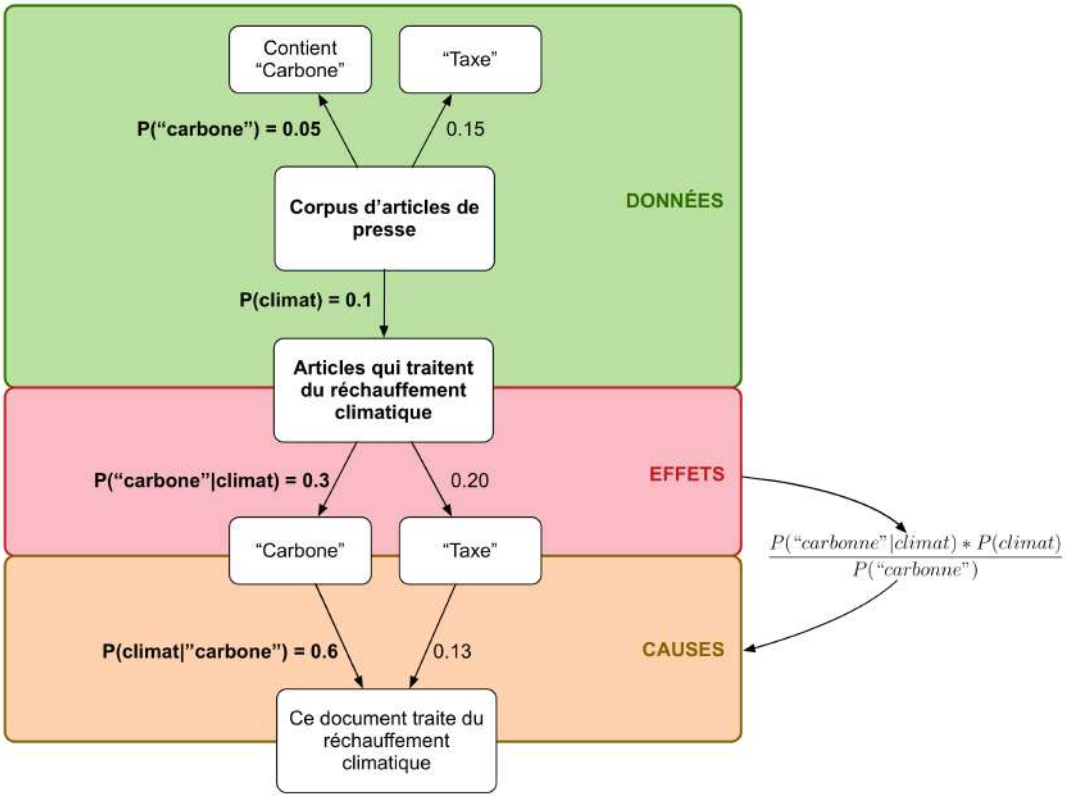


FIGURE 7 – Exemple d'utilisation du théorème de Bayes pour la classification de document.

Afin de construire notre classifieur, c'est à dire un modèle capable d'estimer si un nouvel article de presse parle ou non du climat, il faudrait pouvoir inverser la probabilité des effets observés, c'est à dire pouvoir estimer qu'elle est la probabilité qu'un article porte sur la question du changement climatique sachant qu'il mentionne "carbone", à partir de la probabilité, que nous pouvons calculer depuis les

données, qu'un article mentionne "carbone" sachant qu'il parle du climat (zone orange). Cette "inversion de probabilité", ou calcul de la "probabilité des causes" à partir des effets, est exactement ce que nous permet de faire le théorème de Bayes, via l'équation suivante :

$$P(climat|"carbone") = \frac{P("carbone"|climat) \times P(climat)}{P("carbone")} \quad (2)$$

Cette équation simple nous permet d'estimer la probabilité qu'un nouvel article de presse absent de notre corpus d'apprentissage traite du climat, sachant qu'il contient le mot "carbone". Cependant l'ambition de notre procédure d'apprentissage est de prendre en compte davantage de variables. Si nous ne retenons que deux mots pour illustrer cette procédure d'apprentissage, une stratégie plus réaliste retiendrait probablement tous les mots de tous les articles comme variables, attribuant ainsi à chaque article plusieurs milliers de variables. De plus, certaines méthodes de classification de contenus textuels utilisent la méthode des n-grams qui décompose les mots en retenant toutes les séquences de lettres possible, de toutes les tailles possible entre 1 et n lettres (Bribe de code 2), pouvant constituer ainsi des vecteurs de plusieurs millions de variables. Ainsi, vu le grand nombre de variable dont il est possible de tirer parti, la manière donc celles-ci sont associées constitue le coeur de l'apprentissage avec l'approche bayésienne.

Bribe de code 2 – Un découpage en 5-grams du mot "climat"

```
>>> ngrams("climat", 5)
['^clim', 'clima', 'limat', 'imat*', '^cli', 'clim', 'lima', '
    imat', 'mat*', '^cl', 'cli', 'lim', 'ima', 'mat', 'at*', '^c'
    , 'cl', 'li', 'im', 'ma', 'at', 't*', 'c', 'l', 'i', 'm', 'a'
    , 't']
```

Une des méthodes les plus simples pour combiner ces variables part de l'hypothèse naïve que chaque variable est une cause indépendante qui contribue directement à l'effet qu'on veut prédire. Ce *classifieur bayesien naïf* applatit donc la hierarchie des causes et ignore leurs influences réciproques, au profit de la simplicité de la procédure d'apprentissage et de son implémentation. Cette méthode nous permet donc d'estimer la probabilité qu'un nouvel article de presse traite du réchauffement climatique, sachant qu'il mentionne les mots "carbone" et "taxe", via l'équation suivante :

$$\mathrm{P}(\mathit{climat}|\text{"carbone"},\text{"taxe"}) = \frac{\mathrm{P}(\text{"carbone"}|\mathit{climat}) \times \mathrm{P}(\text{"taxe"}|\mathit{climat}) \times \mathrm{P}(\mathit{climat})}{\mathrm{P}(\text{"carbone"}) \times \mathrm{P}(\text{"taxe"})} \tag{3}$$

L'approche naïve est un classifieur très populaire, du fait de sa simplicité et surtout de son efficacité, malgré son hypothèse presque toujours erronnée de l'indépendance des causes. Cette performance "surprenante" s'explique notamment par le fait que les influences réelles entre les variables ont tendance à s'annuler les unes les autres limitant ainsi le nombre de cas où les ignorer n'est pas pertinent [150]. Par exemple, on peut faire l'hypothèse que dans le cas du classement de document, par exemple en n-grams, il y a une forte redondance entre les variables et que beaucoup d'entre-elles sont peu utiles à la classification, rendant ainsi pertinente l'hypothèse naïve qui annule ces redondances. À minima, ce classifieur bayésien est une application directe du théorème de Bayes, enrichie de l'hypothèse naïve qui permet de traiter plusieurs variables. Pour cela, il est difficile de circonscrire une origine si ce n'est les travaux de Bayes et de Laplace eux-mêmes. Cela étant, on retrouve un premier effort de la part de Duda et Hart [34] pour l'inclure dans la tradition naissante de l'apprentissage en 1973, et une formalisation plus abstraite dans les travaux de Friedman, Geiger et Goldszmidt [37].

Lorsque l'hypothèse naïve est abandonnée, on retrouve alors généralement une structure plus complexe, qu'on nomme *réseau bayésien*. Cette méthode permet de reconstruire une hiérarchie des causes et donc de prendre en compte l'influence des variables entre elles. Par exemple, dans le cas de notre classification d'articles de presse, on pourrait prendre en compte les influences réciproques des mentions de "taxe" et "carbone", c'est à dire la probabilité que "taxe" apparaisse sachant que "carbone" apparait, et/ou inversement. Cependant, un réseau bayésien ne peut pas prendre tous les chemins d'influence en compte, et la structure du réseau, c'est à dire la structure et l'ordre des influences prises en compte, est l'objet de la procédure d'apprentissage. Plusieurs méthodes sont utilisées à cette fin, notamment la recherche d'indépendance entre variables [106], ou l'identification de sous-réseaux possédant le même maximum de vraisemblance [98].

L'usage d'un mécanisme causal dans les réseaux bayésiens est une méthode d'apprentissage artificiel qui s'inscrit par ailleurs dans une réflexion plus ample sur la nature statistique de la causalité. Par exemple, Judea Pearl, un auteur ayant contribué de manière significative à la formalisation de cette famille d'algorithmes, a accompagné

cette recherche de nombreuses tentatives pour distinguer la causalité d'autres types de phénomènes de relations entre variables comme la corrélation et l'influence [94, 95]. Ainsi, si la structure d'un réseau bayésien, c'est à dire la hiérarchie des causes qu'il modélise, peut être un moyen comme un autre de parvenir à un modèle prédictif, il peut, au titre de sa référence aux causes, être une représentation de la réalité qu'il modélise. De cette manière, comme nous l'avons vu pour les arbres de décision, les réseaux bayésiens constituent un mode de représentation des connaissances et des prises de décision pour les systèmes experts. On parle souvent dans ce cas de diagrammes d'influence, particulièrment utilisés dans les années 70 pour l'analyse décisionnelle.

Comme nous allons le voir dans les sections suivantes, plusieurs algorithmes embrassent l'analogie avec l'apprentissage et l'intelligence humaine au-delà de la seule imitation du processus inductif. C'est le processus d'apprentissage lui même qui est imité, par exemple en simulant l'évolution biologique des espèces (§2.3) ou bien le fonctionnement de l'activité neuronale (§2.5). Si ce n'est pas le cas des méthodes bayésiennes, celles-ci sont cependant récemment invoquées par la communauté des neurosciences comme un modèle rendant compte avec justesse de l'apprentissage humain. Stanislas Dehaene affirme ainsi que "l'hypothèse que le cerveau effectue des inférences bayésiennes est l'un des domaines les plus actifs et les plus disputés des neurosciences contemporaines" [31]. Cette hypothèse est notamment explorée par Joshua Tenenbaum pour rendre compte de taches spécifiques comme l'apprentissage du sens des mots par les enfants et les adultes [148], ou comme théorie générale de l'apprentissage humain [132, 133].

2.3 PROGRAMMATION GÉNÉTIQUE

Les algorithmes évolutionnaires ont tendance à ne pas se présenter sous la même hiérarchie de disciplines que les autres procédures d'apprentissages. En effet on parle souvent des algorithmes évolutionnaires, ou de programmation génétique, dans le cadre disciplinaire plus large de l'intelligence computationnelle, c'est à dire des approches où l'inspiration par la nature est particulièrement mise en avant. Sous l'empire de cette classification, chaque méthode est à la fois caractérisée par son inspiration d'un système tiers (biologique, physique) et les problèmes que celle-ci permet de résoudre. Ainsi, les algorithmes de colonies de fourmis résolvent certaines contraintes des systèmes distribués, les systèmes immunitaires artificiels gèrent des formes d'apprentissage et d'accès à la mémoire dans la résolution de problèmes. De la même manière, la programmation génétique si-

mule l'évolution "biologique" de solutions afin de tendre vers leur optimal.

C'est à partir des années 70 que se développe cette approche d'optimisation et de recherche de solution en informatique, fondée sur des "stratégies d'évolution" [107, 120] utilisées notamment pour l'aide au design de l'aérodynamique d'ailes d'avions ou l'étude de la dynamique des fluides. Le livre de John Holland *Adaptation in Natural and Artificial Systems*, publié en 1975 [49], vulgarise ces méthodes et devient la référence la plus communément associée aux origines des algorithmes génétiques. Ce livre illustre son propos dans différents domaines tels que la génétique, l'économie, les jeux, la reconnaissance de motifs, l'inférence statistique, etc. Par ailleurs, Holland avance une preuve formelle qu'un processus d'évolution converge vers de meilleures solutions au fil des générations. Cette définition tournée vers un apprentissage aux domaines d'applications variés annonce clairement l'appartenance de cette famille d'algorithme au *machine learning*. Celle-ci est formalisée comme telle dans le livre de David E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, publié en 1989 [41], où l'auteur consacre un partie importante de son étude à ce qu'il nomme le *Genetic-based machine learning* (GBML), en citant les nombreuses applications de la méthodes dans différents domaines comme : la médecine pour l'établissement d'un diagnostic médical [10], le marketing pour la prédiction de la rentabilité des entreprises [134], la résolution de jeux pour la prédiction des meilleurs paris au Poker [124], le traitement du langage pour la prédiction des séquences de lettres [108], la théorie des jeux pour la résolution du dilemme du prisonnier [8], les sciences politiques pour la prédiction d'évènements internationaux importants [119]. On trouve à partir de la fin des années 80, les premières solutions industrielles basées sur des algorithmes génétiques, notamment produites par General Electric. En 1989, l'entreprise américaine Axcelis commercialise le premier logiciel pour ordinateur personnel, *Evolver*, démocratisant ainsi l'usage des algorithmes génétiques à un plus large public. La médiatisation de ce succès commercial permet de saisir le discours de tels produits. Par exemple en 1990, le *New York Times* fait une revue de ce produit en titrant "survival of the fittest" [72]. Encore aujourd'hui, la sixième version de ce produit est vendue avec une accroche au titre indiquant qu'il fournit "la meilleure solution à n'importe quel problème d'optimisation" [3].

Ce paradigme d'étude des populations a été notamment étendu à la culture et au savoir qui évolueraient selon le même processus. Ainsi, Henry Plotkin [99] définit la connaissance comme le changement produit par elle chez son acquéreur, ouvrant ainsi la perspective d'une analyse évolutionnaire du savoir centrée sur ses détenteurs. Plusieurs

3. http://www.palisade.com/evolver/

auteurs, parmi lesquelles Donald T. Campbell (*in* [103]), offrent une perspective centrée sur la connaissance elle-même, où le principe de vérité de toutes les propositions est testé au fil du temps sur toutes les propositions d'un corpus de connaissances. Pour Popper [100], la falsificabilité d'une connaissance en définit la nature scientifique. Enfin, Richard Dawkins [30] définit le *mème* comme l'atome de connaissance élémentaire : le réplicateur culturel transmis par imitation entre individus à l'instar du gène chez les espèces vivantes. Ces extensions de la théorie de l'évolution de Darwin aux éléments en premier lieu culturels, et non biologiques, de l'activité des espèces, appuient l'hypothèse que les méthodes bio-inspirées de traitement des données peuvent se révéler pertinentes et efficaces à d'autres contextes.

Mais revenons aux les principes structuraux de l'approche évolutionnaire. Le paradigme proposé par Darwin [27] décrit l'évolution dans le temps des espèces vivantes comme le fruit d'une sélection naturelle des individus qui les composent. Chaque membre d'une population est différent et doté de traits particuliers. Plus les *traits* d'un individu sont adaptés à son environnement, plus il est susceptible de se reproduire et donc de reproduire ses particularités au travers de sa progéniture. Ainsi l'espèce évolue en laissant une place plus importante aux traits de ses membres les mieux adaptés. Conceptuellement, une procédure d'apprentissage par algorithme génétique peut se décrire comme suit (cf. Figure 8). la population d'une espèce se compose de solutions possibles, initialement générées au hasard. La pertinence de chacun de ces "individus" est testée sur un jeu de données qui constitue alors l'environnement dans lequel la solution doit "survivre" au mieux. Cette capacité à survivre est mesurée par une fonction de fitness qui détermine le score, ou succès, de chaque solution. On sélectionne alors une certaine proportion des solutions parmi les plus pertinentes et on construit à partir de celles-ci la prochaine génération. Cette dernière est obtenue soit en appliquant de légères modifications aux individus sélectionnés (*mutations*), soit en les combinant (*enjambement*) entre eux. Cette nouvelle génération se voit appliquer la même procédure jusqu'à que la population atteigne un certain fitness, qu'un certain nombre de générations aient été générées, ou bien que la performance des solutions stagne pendant un certain nombre de générations.

L'avantage le plus apparent de ces procédures est qu'elles reposent sur un concept - l'évolution - dont il est facile de se saisir et qui est très intuitif. Il sous-tend bon nombre des explications et analyses sur le changement d'une population par rapport à son environnement, et donc de scénarios d'optimisation et d'apprentissage. Dans ce sens, Alan Turing formule le jeu de l'imitation sous une forme évolutionnaire dans *Computing Machinery and Human Intelligence* :

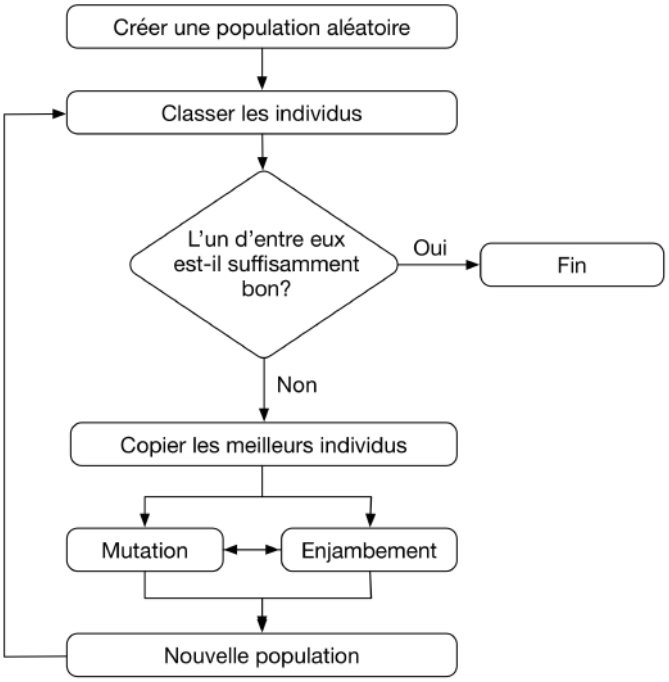


FIGURE 8 – Exemple classique de procédure d'apprentissage par algorithme génétique.

> "Il y a un rapport évident entre ce processus [d'apprentissage] et l'évolution, en identifiant : la structure de la machine comme le matériel héréditaire, ses changements comme des mutations et la sélection naturelle comme le jugement du succès de l'imitation. [...] Puisque il y a probablement un grand nombre de solutions satisfaisantes, la méthode aléatoire semble meilleure que celle systématique. Il faut noter qu'elle est d'ailleurs utilisée dans le processus analogue de l'évolution." [4]
>
> TURING [137]

Il semble ainsi naturel que les processus évolutionnaires aient pu offrir une métaphore idéale et un modèle conceptuel pour décrire une procédure d'apprentissage : l'apprentissage étant alors conçu comme une dialectique alternant changements et adaptation à une contrainte environnementale. En ce sens, le théoricien Leslie Valiant affirme dans son livre de vulgarisation sur la théorie de l'apprentissage automatique :

4. "There is an obvious connection between this process and evolution, by the identifications : Structure of the child machine = hereditary material, Changes of the child machine = mutation, Natural selection = judgment of the experimenter. [...] Since there is probably a very large number of satisfactory solutions the random method seems to be better than the systematic. It should be noticed that it is used in the analogous process of evolution."

> "Les algorithmes d'apprentissage peuvent être exécutés dans des environnements inconnus de leurs artisans, et ils apprennent, en interagissant avec ces environnements, comment agir efficacement en son sein. Après suffisamment d'interactions, ils auront une expertise non pas fournie par l'artisan, mais extraite de l'environnement. [...] J'avance dans ce livre que de tels mécanismes d'apprentissage imposent et déterminent le caractère de la vie sur terre. Le cours de l'évolution est entièrement réalisé par des organismes interagissant avec leurs environnements et s'y adaptant." [5]
>
> VALIANT [138]

La simplicité du schème évolutionnaire le rend aisément transportable à d'autres entités. Il peut être transposé de populations d'êtres vivants à des populations de modèles. Il n'y a pas de limite, et ces modèles peuvent emprunter des formalismes très variés.. En ce sens, l'hypothèse des *building blocks* formulée par Goldberg [41] invite à explorer d'autres "blocs" susceptibles d'être soumis à évolution et recombinaisons au fil des générations. Un exemple assez connu est la régression symbolique. À l'inverse de la régression numérique classique qui cherche à optimiser les paramètres d'un modèle numérique déjà fixé en amont, la régression symbolique détermine la structure du modèle et ses paramètres simultanément [146]. Ainsi, en plus des variables et constantes souvent disponibles dans des approches plus classiques, on trouvera parmi les building blocks des opérateurs mathématiques ($+$, $-$, $\times$, etc) et des fonctions analytiques (cos, sin, tan, etc). Cette approche à eu un retentissement notable dans la communauté scientifique, notamment lorsque des chercheurs sont parvenus à re-découvrir les lois de la mécanique en faisant évoluer des équations entraînées sur les données empiriques de mouvement d'un double pendule [6] [118].

L'approche évolutionnaire s'appuie sur quelques grands principes, laissant un grand nombre de paramètres à la discrétion du modélisateur lors de son implémentation. Il y a bien sur le choix de muter ou d'enjamber les solutions retenues à chaque génération, mais aussi la taille de la population initiale, le pourcentage de la population retenue, la taille des populations suivantes, etc. De plus, la pratique réelle de ce genre de procédure implique bien plus de choix et de marges

5. "Unlike most algorithms, they can be run in environments unknown to the designer, and they learn by interacting with the environment how to act effectively in it. After sufficient interaction they will have expertise not provided by the designer, but extracted from the environment. [...] I argue in this book that such learning mechanisms impose and determine the character of life on Earth. The course of evolution is shaped entirely by organisms interacting with and adapting to their environments."

6. cf. §1.1.1.1

de manœuvre pour "diriger" l'évolution dans une certaine direction, comme l'adapter à une distribution particulière des données, s'adapter aux contraintes de temps et de puissance de calcul disponible ou à une architecture distribuée, augmenter ou diminuer l'espace de recherche, etc. On peut aussi mélanger les procédures de mutation et d'enjambement qui ne sont pas exclusives l'une de l'autre et peuvent être combinées, remplacées ou augmentées par d'autres heuristiques incluant, par exemple, des connaissances du domaine d'application ou de la structure des données. Enfin, il est très facile de construire la fonction de *fitness* pour qu'elle prenne en compte plusieurs objectifs. À ce titre, l'apprentissage multi-objectifs est un domaine où les algorithmes génétiques ont des performances souvent supérieures aux autres algorithmes d'apprentissage. Une hypothèse pour expliquer cette originalité peut être que les algorithmes génétiques permettent d'optimiser assez facilement des situations où l'information est très incomplète ou incertaine et issue de sources multiples.

La malléabilité relative des procédures évolutionnaires pour optimiser un modèle ou une situation constitue aussi une limite de cette famille d'apprentissage. En effet, à mesure qu'on dirige l'évolution dans un certains sens pour anticiper des contraintes ou accélérer la méthode, on peut se retrouver dans une situation où s'en remettre à l'aléatoire des mutations ne représente plus d'intérêt par rapport à une méthode plus systématique. En ce sens, des techniques plus classiques d'optimisation s'avèrent parfois plus efficaces, comme c'est le cas pour le recuit simulé [60] souvent discuté au sein de la communauté des algorithmes génétiques comme une alternative. Pour reprendre l'exemple de la régression symbolique, le succès de travaux de Schmidt et Lipson [118] a donné suite à d'autres perspectives utilisant sur les mêmes blocs des méthodes d'optimisation plus classiques et produisant plus rapidement une solution optimale et déterministe [73].

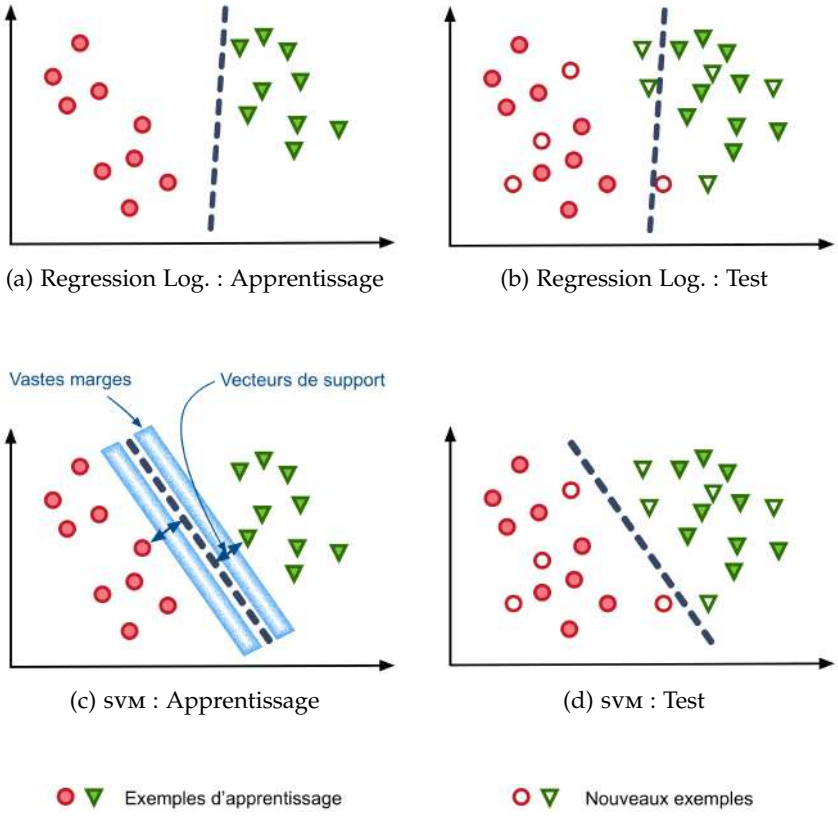
À la différence des autres procédures d'apprentissage exposées ici les algorithmes génétiques reposent sur une analogie qui peut susciter des réactions et des polémiques. Ainsi, si l'histoire politique des théories évolutionnaires peut sembler distincte de son imitation en apprentissage statistique, l'implémentation de telles procédures concernant l'optimisation de situations impliquant l'activité humaine pourrait être justifiée par le schème évolutionnaire et ainsi faire partie de son identié politique. *A minima*, l'hypothèse évolutionnaire présente la première explication scientifique sur l'évolution des espèces en s'appuyant sur le hasard de mutations et leur sélection. À ce titre, elle n'a besoin d'aucune référence à un grand architecte pour fournir sa théorie et oblige ainsi un mouvement créationiste à s'expliquer sur le terrain scientifique, en défendant l'idée que la vie est un objet créé par une ou plusieurs entités et son évolution dirigée par leurs volon-

tés. En ce sens, les théories évolutionnaires sont au centre des conflits entre science et religion depuis le XIX^e siècle, et leur présence dans le programme scolaire fait encore débat aujourd'hui, par exemple aux EUA. Une deuxième polémique qui entoure souvent toute mention de la théorie de l'évolution est celle de son développement en terme de doctrine politique. Ce courant idéologique, le "darwinisme social", considère la survie du plus apte comme une fatalité qui fonde sa nature politique et doit déterminer l'organisation de la société. De manière assez contre-intuitive, on peut voir ce darwinisme social comme une forme de créationisme ayant admis le fait de l'évolution et en l'ayant fait le message de la création devant être suivi. Ce point de vue amplement développé au XIX^e siècle génère bon nombre de théories politiques pronant par exemple la suppression des assistances sociales, des aides pour les pauvres et les handicapés, etc. Ce courant a aussi servi à justifier l'eugénisme et la justification partielle ou totale de l'existence et de hierarchie de races, qui constitue un soutien intellectuel et scientifique au développement des idéologies totalitaires au XX^e siècle. Parmi les auteurs cités comme fondateurs de la statistique dans le chapitre 1, Pearson et Galton ont adhéré aux thèse eugénistes.

2.4 MACHINE À VECTEURS DE SUPPORT

Comme nous l'avons vu dans le chapitre 1 en accompagnant une procédure de régression logistique, apprendre un modèle de classification entre deux groupes revient à trouver une ligne qui les sépare au mieux. Si la métaphore de la ligne reste pertinente dans un cas à deux dimensions comme celui permis par les visualisations en illustration, pour un nombre de dimensions plus important on parle plus volontiers d'hyperplans qu'on peut qualifier plus simplement de surface. Hors c'est bien les questions de classification dans des espaces très importants (avec plusieurs milliers, voir plusieurs millions de dimensions) qui constituent les applications principales des SVM. Nous n'avons pas l'ambition de rentrer dans le détail du formalisme mathématique assez complexe des SVM. C'est pourquoi nous nous contenterons de décrire son principe à l'aide de quelques schémas en deux dimensions.

L'idée, illustrée par la Figure 9, est qu'une classification classique, comme celle permise par une régression logistique, peut produire n'importe quelle ligne qui sépare deux nuages de points distincts. Si les points sont bien séparés, il n'y a plus aucune information supplémentaire quand à la qualité de cette ligne ou modèle. Ainsi, comme on le voit dans la Figure 9a, un tel trait peut très bien se trouver à une distance très inégale des points qu'il classifie correctement par ailleurs. La Figure 9b illustre, dans ce cas, le fait que cette distance inégale tend à favoriser des erreurs lors de la confrontation du modèle



(a) Regression Log. : Apprentissage (b) Regression Log. : Test

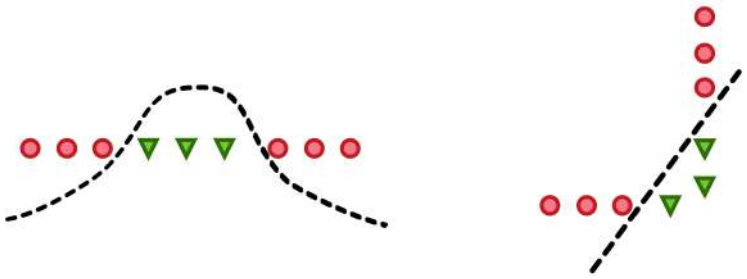
(c) SVM : Apprentissage (d) SVM : Test

FIGURE 9 – Regression logistique Vs. SVM.

à de nouvelles données. En effet, ces nouveaux points auront de plus fortes chances d'être mal catégorisés si les données d'apprentissage étaient déjà proches de la ligne séparatrice. Si nous devions tracer ce modèle à la main, il serait assez naturel de placer la ligne séparatrice à distance égale des deux nuages de points, comme une frontière entre deux territoires probables. Néanmoins, la régression logistique, comme bien d'autres méthodes, ne possède pas d'éléments pour développer cette intuition. C'est à ce niveau là que la méthode des Machines à Vecteurs de Supports intervient. Comme le montre la Figure 9c, les dits "vecteurs de supports" permettent d'établir une distance moyenne entre les points classifiés et le classifieur et ainsi d'optimiser de "vastes marges" qui placent la surface du modèle à distance égale et maximale des points les plus proches de part et d'autre de la frontière. Ainsi, comme illustré dans la Figure 9d, les nouveaux exemples sont moins fréquemment susceptibles d'être mal catégorisés même à proximité de la frontière. Les premières formulations de cette idée ont été publiées de manière indépendante dans les travaux de Vapnik et Lerner en 1963 [139], et de Duda et Hart en 1973[34]. Vapnik et Chervonenkis [141] ont par la suite proposé une formulation mathématique plus rigoureuse.

Mais l'originalité fondamentale des SVM tient sans doute à leur capacité à identifier des solutions non linéaires. L'exemple que nous avons décrit est linéaire au sens où la droite séparant deux sous espaces est bien linéaire. Comme nous l'avons vu dans le chapitre 1, la régression logistique peut produire des modèles simples, multipliant chaque variable par un poids, définissant un trait droit qui représente le modèle linéaire de classification. Cependant, afin de permettre à ce trait d'épouser des formes plus complexes, comme illustré dans la Figure 10a, on peut décider de complexifier le modèle pour permettre des découpages plus complexes de l'espace des variables, typiquement en multipliant les variables entre elles, en les mettant au carré, au cube, etc. Les SVM proposent une autre approche à ce problème, connue sous le nom "d'astuce du noyau" et communément attribuée aux travaux de Aizerman et al. [5]. L'idée est qu'au lieu de complexifier le modèle, on peut rajouter des dimensions aux données afin qu'un modèle linéaire puisse les séparer. Cette astuce est illustrée par la Figure 10b, où les données sont réarrangées de sorte qu'une ligne droite puisse séparer deux familles de points, ce qui aurait été impossible dans l'espace initial avec des opérateurs purement linéaires.

Les vastes marges et l'astuce du noyau constituent donc les véritables innovations introduites par les SVM. Bien que toutes deux datent des années 60, ce n'est qu'au début des années 90 que plusieurs chercheurs, dont Boser et Vapnik, les combinent pour donner naissance à la formulation contemporaine de cette méthode [15] qui ne sera largement adoptée qu'après 1995 lorsque Vladimir Vapnik publie un livre



(a) Modèle non-linéaire (b) Modèle linéaire avec noyau

FIGURE 10 – Illustration de l'astuce du noyau.

[140] qui inscrit cette méthode comme le prolongement naturelle de *la nature de la théorie de l'apprentissage statistique*. En effet, Vladimir Vapnik, auteur central dans l'élaboration d'un formalisme adapté aux SVM et ce depuis les années 60, propose dans ce livre pédagogique de les situer comme un juste retour aux origines de l'apprentissage et une alternative aux réseaux de neurones. Outre cette perspective historique quant à l'apprentissage, il situe les SVM dans un débat philosophique sur la nature de l'intelligence humaine :

> "Le problème d'estimer les valeurs d'un fonction à un moment donné traite une question qui est discutée en philosophie depuis plus de 2,000 ans. Quelle est le fondement de l'intelligence humaine : la connaissance des lois (règles) ou la culture d'un accès direct à la vérité (intuition, inférence *adhoc*) ?" [7]
>
> VAPNIK [140]

Par ces mots il reformule le débat déja commenté par Breiman [20] entre interprétabilité et performance d'un algorithme, en l'inscrivant dans l'analogie avec l'intelligence humaine. Doit-on imiter une intelligence qui rend compte des règles, qui est interprétable, ou bien une intelligence qui, de manière plus difficile à interpréter, développe les intuitions les plus justes ? Les SVM favorisent l'intuition et remettent en cause une opinion commune au sein de la communauté statistique qui favorise l'idée de modèle explicite quant à ses règles. Ainsi, comme Breiman pour les forêts aléatoires, Vapnik accompagne la formalisation et l'explication de sa méthode d'apprentissage d'une justification, ici épistémologique plus que d'usage, d'un nouveau compromis entre interprétabilité et précision.

7. "The problem of estimating the values of a function at a given point addresses a question that has been discussed in philosophy for more than 2000 years : What is the basis of human intelligence : knowledge of laws (rules) or the culture of direct access to the truth (intuition, adhoc inference) ?"

Les qualités principales des SVM sont nombreuses. Tout d'abord, cette méthode tend à mieux généraliser, c'est à dire à être moins prisonnière des détails des données, ou de mieux interpréter et isoler les motifs qui seront pertinents pour de nouvelles données. Ensuite, les SVM ont plusieurs propriétés mathématiques qui en font une procédure d'apprentissage qui garantit de trouver un optimum global, c'est à dire la meilleure solution étant donnés les exemples disponible pour l'apprentissage. Les SVM convergent rapidement vers cette solution et peuvent donc être utilisées sur des données décrites par de nombreuses variables. Enfin, la complexité mathématique de la procédure n'empêche pas une utilisation simple via les librairies de programmation comme LIBSVM. En effet, de la même manière qu'on n'implémente pas soi-même une fonction de log, la complexité de la procédure envisagée invite son utilisateur à l'utiliser telle quelle et de tirer avantage qu'elle ne nécessite que très peu de paramètres pour être mise en œuvre de façon générique.

Ces nombreux avantages en ont fait un outil cohérent, efficace et facile d'utilisation qui, depuis sa formulation dans les années 90, est un terrain de recherche actif et une méthode utilisée dans des communautés académiques variées. Cependant le compromis entre interprétabilité et efficacité qui fonde cette méthode ne semble pas faire l'unanimité dans ses usages et on retrouve plusieurs travaux qui essayent de reconstruire les règles du modèle issues de la procédure d'apprentissage par SVM, afin de pouvoir profiter de cette performance accrue quand une solution plus explicite et transparente est souhaitée. On trouve de tels travaux notamment en biologie pour la classification de séquences génétiques [105].

2.5 RÉSEAU DE NEURONES ARTIFICIELS

Les réseaux de neurones ont des origines fortement ancrées dans les neurosciences, et plus précisément dans la modélisation des interactions entre neurones dans le cerveau. La première formulation de l'activité d'un neurone artificiel, le perceptron, est d'ailleurs le fruit d'une collaboration entre un psychiatre, Warren McCulloch, et un mathématicien, Walter Pitts [75]. Ce travail publié en 1943 prétend avant tout fournir un outil d'expérimentation en neurophysiologie. Partant de l'observation que l'état d'un neurone est binaire, c'est à dire soit actif, soit passif, les auteurs explorent la possibilité que le fonctionnement de l'activité neuronale puisse être décrit par un raisonnement logique. L'idée, illustrée par la Figure 11, est que le perceptron reçoit un ensemble de signaux à son entrée, qui vont être pondérés au sein du réseau de neurones. Si la somme de toutes ces valeurs dépasse un certain seuil alors le neurone s'active (il produit 1), sinon il reste éteint (il produit 0).

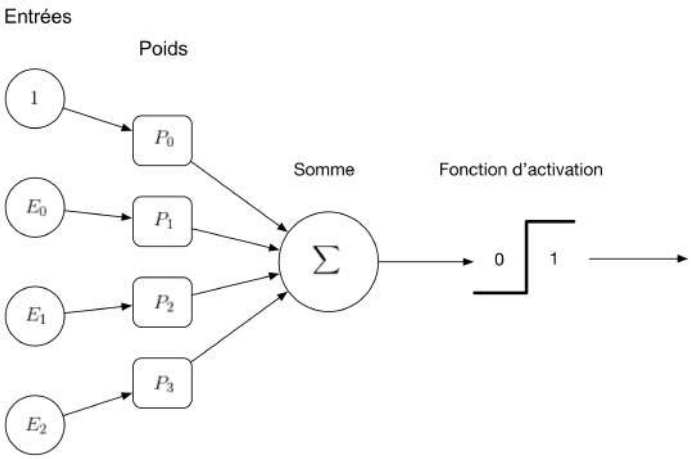


FIGURE 11 – Le perceptron (neurone artificiel à seuil binaire).

Pour transformer ce modèle de neurone en dispositif d'apprentissage se pose une des questions centrales à la recherche en réseau de neurones : comment saisir l'erreur du comportement d'un neurone et ajuster le dispositif en fonction de celle-ci ? Dans le cas du modèle de McCulloch et Pitts, le problème est résolu en soustrayant ou en rajoutant aux poids la valeur de leurs variables respectives, selon que le neurone, s'active ou non, à tort ou à raison. Selon cette procédure, il est garanti que les poids corrects seront trouvés, à condition que les variables rendent suffisamment compte du problème que l'on souhaite résoudre. Ce modèle est donc fortement limité quant à ce qu'il est capable d'apprendre, mais il est néanmoins encore utilisé aujourd'hui pour certaines application nécessitant un système simple pour résoudre un problème avec trop de variables pour que les algorithmes plus complexes soit utilisés.

Le perceptron est popularisée par Rosenblatt qui, en 1957 [110], l'introduit comme un dispositif d'apprentissage, pour ensuite, en 1958 [111] et 1961 [112], s'écarter de l'apprentissage artificiel et utiliser le perceptron comme un moyen d'expérimentation afin d'enrichir une théorie du cerveau et de ses dynamiques. Le succès de cette approche et son soutien financier par la marine militaire américaine attire l'attention de la presse qui met l'accent sur les vertus d'automation du modèle : "L'embryon d'un ordinateur électronique dont la Marine espère qu'il marche, parle, voit, écrive, se reproduise lui-même et soit conscient de son existence" [8] (*The New York Times*), "Le monstre Frankenstein construit par un robot de la Marine qui pense" [9] (*Oklahoma*

8. "The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence. Later perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech and writing in another language, it was predicted."

9. "Frankenstein monster designed by the Navy Robots that Thinks"

Times). La métaphore du neurone associée à une capacité d'apprentissage constituent un terrain propice à la représentation médiatique "fantasmée" d'une machine autonome, évolutive, qui par défaut imite l'homme et menace de le "dépasser". L'ampleur rapide que prend le perceptron de Rosenblatt attise également les critiques, notamment celles formulées par Minsky et Papert [79], qui pointent vers certaines limites du modèle, notamment son incapacité à apprendre des mouvements logiques simple comme la fonction XOR (ou-exlusif) [10]. Cette critique est largement reprise par les défenseurs de l'approche symbolique de l'IA, qui dirigent leurs efforts vers une représentation formelle des problèmes, et une résolution de ceux-ci par la logique et la recherche d'une solution optimale. Ce courant de l'IA parvient alors à écarter la recherche sur le perceptron et les réseaux de neurones des financements publics et ce n'est qu'au début des années 80, lorsque la puissance de calcul devient plus accessible, que celle-ci reprend [91].

Pour obtenir de meilleurs résultats, ce n'est pas tant le comportement du neurone, ou le nombre de neurones, qui est modifié, mais la mise en place de structures plus complexes où les neurones sont ordonnés au sein de différentes couches interconnectées. L'idée est de rajouter une ou plusieurs couches intermédiaires, ou "cachées", entre le résultat final à prédire et celle des variables, afin de permettre à davantage de combinaisons d'être explorées. La Figure 12 montre comment un tel réseau de neurones n'est en fait qu'un perceptron multi-couches, où chaque unité réalise la même opération de s'activer ou non, en fonction d'entrées qui, elles, diffèrent, d'une unité à une autre.

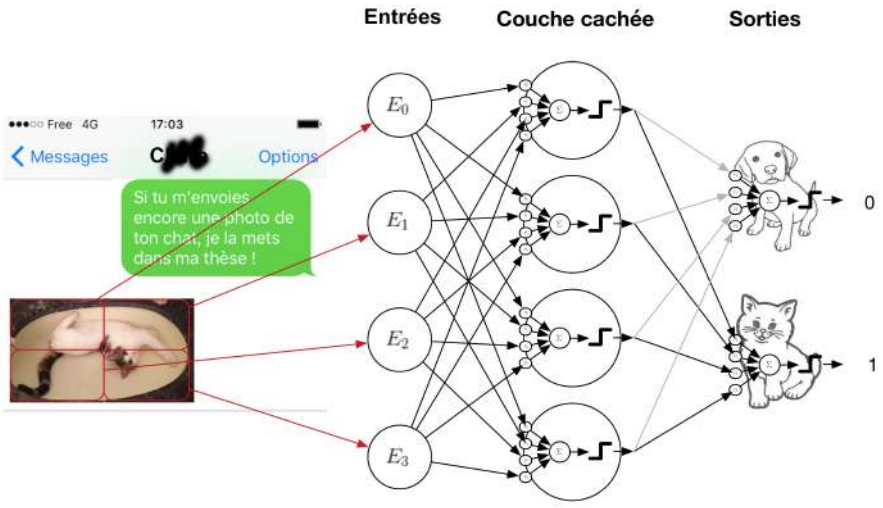


FIGURE 12 – Un réseau de neurone avec une couche cachée (*hidden layer*).

L'idée découle de la limitation inhérente au perceptron : son succès pour apprendre une tâche dépend largement des variables qui lui

10. Si cette critique est juste pour le perceptron unique, elle s'est avérée erronnée lorsqu'on met un nombre suffisant de neurones en réseau.

sont fournies en entrées. Il s'agit en fait d'un problème transverse à tous les algorithmes d'apprentissage (arbre de décision, SVM, etc.) : sélectionner des variables pertinentes pour un problème donné, et les ajuster en les combinant les unes avec les autres. Par exemple, la largeur et la hauteur d'un objet peuvent être des caractéristiques de celui-ci que l'on introduit en données d'entrées. Mais ces données seront insuffisantes si la propriété réellement critique décrivant le système est le rapport entre les deux ($\frac{largeur}{hauteur}$). À ce titre, Domingos [32] identifie l'ingénierie des variables (*features engineering*) comme un élément clé de l'apprentissage artificiel. Pour palier à ce problème et tirer avantage de la structure modulable d'un réseau de neurones, plusieurs chercheurs vont tenter de déléguer la construction des variables au réseau de neurones lui-même. L'idée est qu'au lieu de construire des variables en amont du réseau, il faut donner à celui-ci les données sous leur forme la plus élémentaire (le pixel, la lettre), sans combinaisons ou filtre, et laisser les premières couches construire les variables dont le réseau a le plus besoin pour l'étape de classification des couches suivantes. Le fait qu'un dispositif d'apprentissage puisse assumer une fonction de classifieur, mais, aussi, de construction des variables de celui-ci, est caractéristique de ce que l'on appelle l'apprentissage profond (*deep learning*). Ces réseaux profonds, ont pour objectif d'inclure dans le réseau, la classification et la construction des variables optimales pour celle-ci, mais aussi la préparation des données, leur post-traitement, etc [70]. À minima, chaque neurone d'une couche d'un réseau est une combinaison des variables de la couche précédente. Ainsi, on rajoute des couches, ou calques, à un réseau qui "contraint" celui-ci à tantôt déceler des motifs pertinents et donc réduire le nombre d'inputs (*convolution*), tantôt à augmenter le nombre d'inputs afin d'accroitre le nombre de combinaisons observées, afin de juger si elles sont pertinentes (*sub-sampling*). La Figure 13 permet d'apprécier comment ce processus est réalisé plusieurs fois dans un même réseau visant à reconnaitre des caractères manuscrits[69].

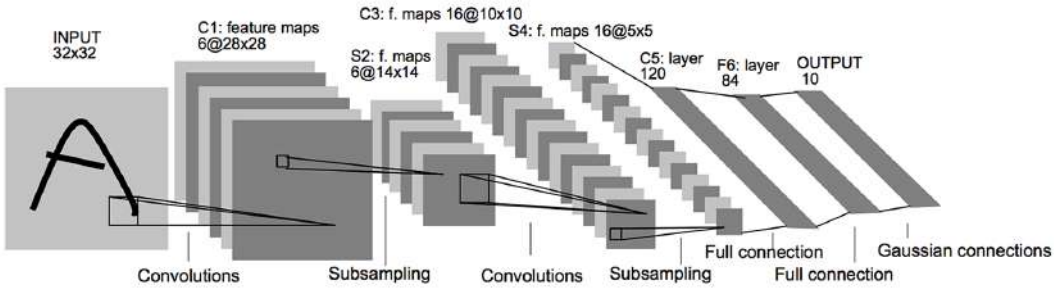


FIGURE 13 – Architecture du réseau convolutionnel *LeNet-5* pour la reconnaissance de caractères [69].

Le comportement d'une seule unité ou d'une couche du réseau est assez prévisible et répond à des règles bien établies relatives à la fonction d'activation du neurone, le nombre d'unités par calque, etc. Néanmoins la structure du réseau dépend nettement plus des contraintes matérielles et du temps d'apprentissage envisageable. Pour cela, étant donné un objectif et une base de données d'entrainement, trouver la bonne architecture relève en partie de l'intuition de l'auteur du réseau, et ne peut être qualifiée d'optimale que relativement à d'autres architectures moins efficaces. Cette responsabilité de l'"artisan" d'un réseau quand à son architecture entraine une personnification de celle-ci, et ils sont parfois baptisés du nom de leur auteur : *LeNet* pour le travail de LeCun sur la reconnaissance de caractères [68], ou *AlexNet* pour le réseau de Alexandre Krizhevsky sur le reconnaissance d'objets [61]. C'est une des raisons pour lesquelles on ne retrouve pas les réseaux de neurones dans les librairies de programmation généralistes d'apprentissage, car alors que la plupart des autres algorithmes ne nécessitent pas ou peu de paramètres, un réseau de neurone repose sur une architecture qui dépend fortement du problème qu'il doit résoudre [11].

Un des goulots d'étranglement majeurs lors de la reprise de la recherche sur les réseaux de neurones artificiels a été, pour ces réseaux plus complexes, de réussir à rétro-propager les erreurs de prédiction en terminaison du réseau. C'est à dire comment - pour chaque exemple connu et présenté au réseau - réajuster les poids du réseau pour s'approcher d'une meilleure solution. Si ce problème était résolu de manière assez simple pour le perceptron, les architectures complexes empêchent une solution équivalente et le calcul de l'erreur est impossible à réaliser de la même manière. Plusieurs travaux traitent de ce problème dans le contexte des réseaux de neurones [66, 145] mais c'est le travail de Rumelhart et al. en 1988 [114] qui donnera sa cohérence contemporaine à l'algorithme, nommé rétro-propagation du gradient (*backprop*). Bien que l'algorithme ne soit pas très intuitif et comporte des subtitlités mathématiques complexes, Geoffrey Hinton, l'un de ses auteurs, n'hésite pas à le simplifier de la sorte : au lieu d'essayer de mesurer l'erreur proprement dite, la rétropropagation du gradient se concentre sur la vitesse d'évolution de l'erreur [12] [47]. Une fois cette mesure faite sur la dernière couche du réseau, celle qui instancie la prédiction, l'algorithme de *backprop* modifie les poids qui lient cette dernière avec la précedénte et répète la mesure et les corrections des couches précédentes, jusqu'à parvenir à la première couche directement connectée aux données d'entrée et avoir modifié tous les poids du réseau afin que l'erreur perçue diminue.

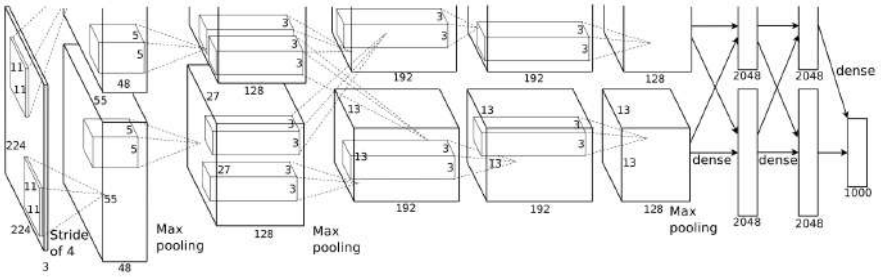
11. Ces aspects sont abordés de manière plus comparative dans la section §2.6

12. "The idea behind backpropagation is that we don't know what the hidden units ought to do, but we can compute how fast the error changes as we change a hidden activity."

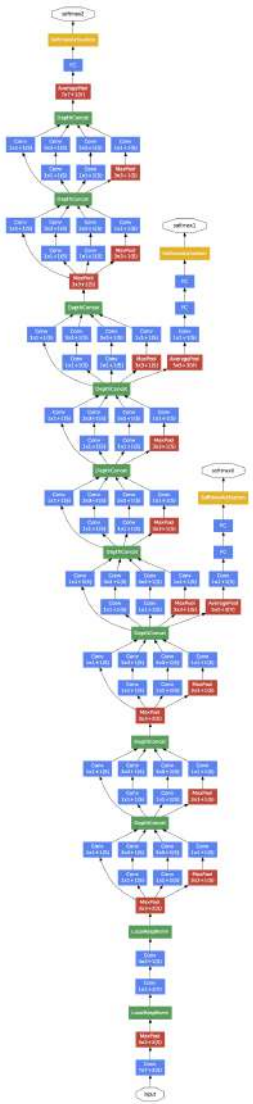
La complexité du calcul de l'erreur via *backprop* augmente considérablement avec la taille de l'architecture du réseau (nombre de couches, nombre de neurones par couche), et on doit en grande partie le succès récent du *deep learning* à une implémentation distribuée lors de la compétition de reconnaissance d'objets *ImageNet*. En effet, lors de l'édition de 2012 de cette compétition, Krizhevsky et al. [61] utilisent des cartes graphiques (GPU) destinées au rendu des graphismes de jeux qui permettent à des milliers d'opérations simples d'êtres exécutées simultanément. Cette technique permet d'entrainer un réseau de neurones traitant les millions de variables d'une image haute résolution sur de nombreuses couches. En 2012, cette équipe remporte la compétition avec une réseau de 9 couches, en 2014 une équipe de Google la remporte avec un réseau de 40 couches, et en 2015 Microsoft gagne avec un réseau de 152 couches (Figure 14). Le nombre accru de couches et l'utilisation de GPU sont devenus en quelques années à peine la méthode privilégiee pour implémenter des réseaux de neurones, et constituent un élément important du succès contemporain de l'apprentissage profond.

On voit bien, donc, que les réseaux de neurones on su sortir d'un "hiver" à deux reprises, une première fois dans les années 80 grâce à la backprop qui a permis la calcul de réseaux multicouches pour désarsonner la polémique du perceptron, et une deuxième fois dans les années 2000 grâce à l'implémentation distribuée sur des cartes graphiques. Cela constitue un cas relativement emblématique de l'histoire récente de l'apprentissage artificiel, et de comment, en aval d'une métaphore et d'un axiome général, viennent contribuer des éléments très pragmatiques d'algorithmie, de matériel, de coût financier et d'espace et de temps de calcul, etc.

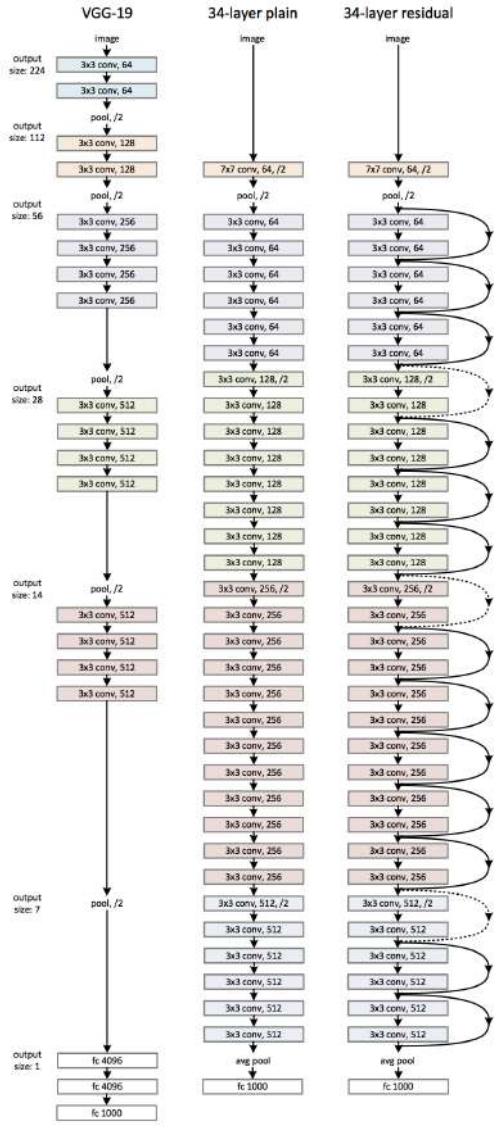
Malgré leur performance, les réseaux de neurones possédant de multiples couches intermédiaires entre leurs entrées et leurs sorties sont délaissés par de nombreuses applications industrielles de prédictions et de classifications car les modèles qu'il produisent apparaissent comme ininterprétables. De plus, la procédure d'apprentissage comporte de nombreux minima locaux qui font que plusieurs modèles utilisent les variables de manières différentes à performance équivalente. Les variables en entrée sont combinées récursivement à de nombreuses reprises et il semble impossible de reconstruire, et donc d'expliquer, lesquelles ont une importance particulière, ou bien des relations de causalité avec les sorties. Une première solution, souvent ignorée, sont les tests de sensilité (*sensitivity analysis*) qui permettent *a posteriori* de saisir le rôle des variables d'entrée dans le fonctionnement du réseau [29, 52, 71]. Une piste d'avantage explorée pour augmenter l'intelligibilité d'un modèle dans le cas de la reconnaissance d'image est la description, voire la visualisation des couches intermédiaires du réseau. Si le réseau assemble les pixels d'une image de



(a) 9 couches de *AlexNet* [61]



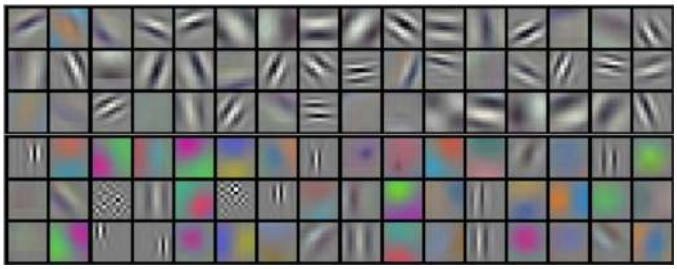
(b) 40 couches de *GoogLeNet* [130]



(c) 152 couches par Microsoft [46]

FIGURE 14 – Architectures des réseaux de neurones des équipes ayant remportées les compétitions *ImageNet* en 2012 (a), 2014 (b) et 2015 (c).

couche en couche afin de construire les variables pertinentes à la classification, en visualisant ces étapes intermédiaires on peut observer quelles formes font sens lors de la procédure d'apprentissage. Ainsi, la Figure 15a montre que des éléments premiers de formes (ligne, contours) se dégagent au sein des premières couches du réseau, et la Figure 15b et la Figure 15c montrent comment le réseau abstrait une représentation minimale des catégories visées dans la dernière couche du réseau.



(a) Formes dans les premières couches [61]

(b) Visage d'un chat [65]

(c) Visage d'un humain [65]

FIGURE 15 – Visualisation de couches intermédiaires d'un réseau de neurones.

2.6 TYPOLOGIES ET ANALYSES COMMUNES

Dans les sections précédentes nous venons de parcourir les principaux algorithmes d'apprentissage qui constituent ce domaine de connaissance. Cette liste n'est pas exhaustive et ne prétend pas épuiser toutes les pistes de recherche explorées pour "résoudre" le problème de l'apprentissage. Comme mentionné en introduction de ce chapitre, la liste des algorithmes que nous venons de détailler est très proche de celle qui constitute la typologie de Domingos dans *The Master Algorithm* [33]. Domingos traite des communautés que nous avons présentées dans ce chapitre comme des "tribus". Il file la métaphore de cette dénomination et parle ainsi des "territoires" de chacune, de leurs "frontières" communes, leurs "conflits", etc. Le Tableau 5 permet d'observer les principaux éléments qui caractérisent chacune de ces tribus.

"Tribu"	Inspirations	Algo. fondamental	Algo. appliqués
Symbolistes	philosophie, psychologie, logique	déduction inverse	arbre de décision, forêt aléatoire
Connectionnistes	rétro-ingénierie du cerveau, neurosciences, physique	rétropropagation du gradient	réseaux convolutionnels, récurrent
Évolutionnistes	évolution, génétique, biologie	programmation génétique	mutation, enjambement
Bayésiens	probabilités, statistiques	inférence bayésienne	classification naif bayésienne, modèles graphiques probabilistes
Analogistes	psychologie, optimisation mathématique	SVM	Noyaux

Tableau 5 – Typologie des *tribus* de l'apprentissage artificiel selon Domingos [33].

INSPIRATIONS ET SIMULATIONS

Cette typologie nous permet de diviser les tribus de l'apprentissage en fonction de différents critères, notamment leurs sources d'inspirations respectives. Dans le chapitre 1 nous avons vu comment l'inspiration des processus d'apprentissage chez l'humain et l'animal était fondatrice de l'effort de le reproduire chez la machine. Dans le présent chapitre, nous avons vu que les réseaux de neurones et les algorithmes génétiques poursuivent cette stratégie d'imitation avec une partie de leurs acteurs qui assurent que c'est l'imitation des moyens (le cerveau, l'évolution) qui serviront le mieux l'imitation de la fin (l'apprentissage). De cette manière, si l'axiome est l'apprentissage, il faut imiter ce que l'on sait qui apprend de l'expérience, comme le cerveau et l'évolution biologique. Cela constitue une inspiration pour une partie des chercheurs en réseaux de neurones, et en programmation génétique. Ces inspirations nous permettent une première division au sein des familles d'algorithmes mentionnées, entre celles qui sont bio-inspirées, et celles qui ne le sont pas.

Domingos, dans sa typologie, complète les inspirations que nous n'avons pas mentionnées pour les autres algorithmes traités ici, notamment en empruntant à la psychologie, la philosophie ou la logique. De manière générale, il apparait que chaque algorithme est ancré dans une métaphore à la fois plus grande que lui, mais toujours plus spécifique qu'une référence à l'apprentissage au sens large. On pourrait alors séparer les algorithmes qui s'inspirent du vivant, de ceux qui s'inspirent de traditions scientifiques (causalité, déduction inverse, etc), ou plus simplement, faire comme Domingos, et séparer chaque type d'algorithme au sein de sa tribu respective.

Si ces inspirations portent une origine, on a vu aussi dans ce chapitre des analogies qui jouent un rôle similaire d'inspiration mais *a posteriori*. C'est la cas notamment des modèles bayésiens qui, bien qu'ils ne soient pas bio-inspirés, inspirent des théories contemporaines de l'apprentissage humain en neurosciences. Ici, donc, c'est le modèle d'apprentissage artificiel qui inspire un modèle d'interprétation du vivant et permet, même dans un retour et dans un second temps, de faire du fonctionnement du cerveau une source d'inspiration pour les réseaux bayésiens.

Enfin, on a pu voir que les algorithmes explicitement bio-inspirés, comme les réseaux de neurones et les algorithmes génétiques, ont des origines non seulement dans l'imitation du vivant mais aussi dans sa simulation. Leurs premières formulations étaient proches de travaux cherchant à simuler les processus vivants pour mieux pouvoir les étudier. Si l'effort de l'apprentissage artificiel est bien distinct de celui de la simulation, il ne faut pas sous-estimer le dialogue qu'il peut y avoir entre ces deux approches, ou du moins un terrain de connaissance commune. Plus concrètement, une procédure à fin de simulation pourra tout à fait délaisser son ambition d'optimisation au profit de la représentativité du modèle avec les phénomènes étudiés. Inversement, une procédure d'apprentissage est mise en place pour optimiser la résolution d'un problème et la cohérence du modèle avec le phénomène de construction dont il s'inspire n'a pas sa place comme contrainte à cet effort. Ces deux approches ont donc des objectifs bien différents mais partagent les formalismes fondamentaux qui permettent leur mise en place respective.

IMPLÉMENTATIONS ET USAGES GÉNÉRIQUES

Dans un premier temps nous avons vu que certains algorithmes, comme les SVM, offrent un usage relativement générique de leurs procédures d'apprentissage. De fait, on peut opérer cette procédure sur un jeu de données sans avoir à décider de nombreux paramètres ni à implémenter la procédure d'apprentissage elle-même. Il en est de même pour plusieurs des algorithmes discutés ici, plus précisément les arbres de décision, les forêts aléatoires, les classifieurs naïfs bayésiens et le perceptron. Le caractère générique de ces procédures n'empêche pas qu'elles puissent être raffinées par plusieurs paramètres que nous ne détaillons pas ici, mais il permet surtout à ces méthodes d'être largement diffusées et faciles d'utilisation même par un néophyte. Ainsi, la Bribe de code 3 montre comment on peut en 20 lignes de code informatique à peine, solliciter 5 des classifieurs les plus génériques sans avoir à faire autre chose que les nommer. Cette capacité d'un algorithme à pouvoir être appelé et utilisé de manière générique est un élément décisif quant à son adoption au-delà de sa communauté de recherche car cela permet à de nombreuses personnes incapables

de saisir la complexité mathématique de ces modèles, ou de les implémenter, de néanmoins profiter de leurs vertus prédictives, voire de les mettre en compétitions les uns avec les autres.

Bribe de code 3 – Comparaison de 5 classifieurs génériques sur les données *Iris*

```
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
from sklearn import naive_bayes
from sklearn.linear_model import Perceptron
from sklearn import svm

from sklearn import datasets
from sklearn.cross_validation import train_test_split
iris = datasets.load_iris()
X_train, X_test, y_train, y_test = train_test_split(iris.data,
    iris.target, test_size=0.33)

for classifier, name in [
   (tree.DecisionTreeClassifier(), "Decision Tree"),
   (RandomForestClassifier(), "Random Forest"),
   (naive_bayes.GaussianNB(), "Naive Bayes"),
   (Perceptron(), "Perceptron"),
   (svm.SVC(), "SVM")]:

   score = classifier.fit(X_train, y_train).score(X_test, y_test)
   print name, score
```

Bribe de code 4 – Résultat de Bribe de code 3

```
$ python code/exemple_sklearn.py
Decision Tree 0.94
Random Forest 0.94
Naive Bayes 0.94
Perceptron 0.72
SVM 0.96
```

Ici, leurs fonctions respectives reposent sur une implémentation en amont faite par une librarie de programmation populaire dans ce domaine [13]. Une telle implémentation est beaucoup plus difficile pour des procédures dont une abstraction générique est plus complexe à extraire comme c'est le cas pour les réseaux de neurones et les algorithmes génétiques. Comme nous l'avons vu lors de la présentation de ces algorithmes, ceux-ci reposent sur de nombreux critères (architecture du réseau, fonction d'activation, population, fonction de fitness) qui dépendent fortement des données traitées, de l'objectif

13. Sci-kit Learn : http://scikit-learn.org/.

que l'on cherche à optimiser, des contraintes du calcul (temps, espace mémoire). Il existe bien sur des implémentations génériques de ces procédures, mais elles ne permettent que très peu d'utiliser les performances des algorithmes sous-jacents [14]. Il s'agit donc d'un nouvelle dimension à prendre en compte dans le reste de notre étude : la capacité d'un algorithme à devenir un dispositif exportable de manière générique hors de son champ de recherche, pour expliquer la diffusion de son usage et la multiplication de ses champs d'application.

INTERPRÉTABILITÉ, COMMUNICABILITÉ ET RÉTRO-INGÉNIERIE DES SYSTÈMES D'APPRENTISSAGE

Comme nous l'avons vu dans le chapitre précédent, une partie des dissensions internes à l'IA pendant la seconde moitié du XX^{e} siècle s'exprimaient en opposant approche symboliste à l'approche connexionniste. L'approche symboliste, comme elle s'astreint à des représentations intelligibles des données en jeu permettait une certaine forme de contrôle ou de compréhension des décisions prises par l'IA. Ensuite, les travaux de Breiman [20], nous ont permis de formuler un compromis entre interprétabilité et précision comme le pivot d'une nouvelle culture statistique qui a fait le pont entre une approche traditionnelle et une approche algorithmique au service des objectifs de l'IA.

Tout au long du présent chapitre nous avons mentionné et expliqué en quoi les différents algorithmes convergeaient vers des solutions plus ou moins interprétables. Ainsi, on peut classer chacun d'eux selon la typologie de Breiman et affirmer que les approches symbolistes et bayésiennes sont traditionnelles et les autres algorithmiques. Si Breiman avait formulé son propos pour justifier le fait que les forêts aléatoires délaissaient la description au profit de la précision, nous avons vu qu'un mouvement similaire a accompagné la formulation des SVM par Vapnik qui justifie la même position dans ce compromis par des références épistémologiques pour écarter la nécessité de règles au profit d'intuitions moins interprétables.

Galit Shmueli [121] propose un complément à cette typologie binaire et affirme qu'il serait plus juste de distinguer entre les modèles : ceux qui font appel à la notion de causalité et donc expliquent (les réseaux bayésiens), ceux qui permettent de décrire (arbres de décision), et ceux qui favorisent la performance de la prédiction (SVM, réseaux de neurones et algorithmes génétiques). Qu'il s'agisse d'explication (causalité, bayésien) ou de description (induction, arbre de décision), nous avons pu observer que les modèles interprétables comportent certaines failles quant à leur cohérence ou à l'universalité de leur pro-

14. Dans ce sens, la version de développement de sci-kit learn (v.0.18) comprend quelques modèles simplifiés de réseaux de neurones utilisable de manière générique.

pos, principalement car un même modèle ordonnant les variables de manière totalement différente pourrait produire les mêmes prédictions, et ainsi prouver que la première explication n'en est qu'une possible parmi tant d'autres. Ainsi, un modèle traditionnel ne fournit pas une explication, une vérité, mais une histoire possible qui justifie son modèle et c'est parfois plus pour la possibilité de partager, communiquer, commenter le modèle qu'on privilégiera ce type d'approche. On peut donc réduire en grande partie l'interprétabilité à la communicabilité.

Comme nous l'avons vu pour les réseaux de neurones, des méthodes existent pour, malgré l'obfuscation de la procédure d'apprentissage, recontruire l'importance des variables dans l'établissement des prédictions, notamment par des tests de sensibilité. Ce genre de méthodes permet donc de satisfaire partiellement l'impératif de communicabilité du modèle, par une forme d'explication non-universelle, non-exhaustive. Ainsi, la différence faite communément entre les algorithmes interprétables et ceux qui ne le sont pas n'est pas forcément pertinente d'un point de vue épistémologique, bien qu'elle semble très pertinente pour expliquer des usages différents dans des communautés qui associent à certaines procédures une intelligibilité "naturelle".

Enfin, la mise en place d'un système d'apprentissage à grande échelle s'écarte souvent de la pratique académique d'un algorithme unique sur un jeu de données précis. Il s'agit plutôt d'une association complexe de procédures variées, connectées les unes aux autres tant dans leurs algorithmes et leurs paramètres que du point de vue de leurs jeux de données, souvent hétérogènes. Cela peut favoriser donc des méthodes plus globales, plus proches d'une forme de rétro-ingénierie qui considère le système d'apprentissage comme une boite noire et ne s'appuie que sur l'observation de ses entrées et sorties pour établir une explication *possible* de son comportement.

RÉSUMÉ DU CHAPITRE 2

Dans ce chapitre nous avons décrit plusieurs des principales procédures d'apprentissage artificiel. Il s'agit de véritable *épistémès* distinctes les unes des autres, portant leurs efforts d'apprendre de l'expérience par des inspirations, des formalismes et des structures de données qui façonnent leurs contraintes, coûts, bénéfices, capacités à être socialisées via l'interprétation de leur modèle, etc.

Les arbres de décisions s'inspirent de la représentation de systèmes experts de prise de décision et empruntent des méthodes à la théorie de l'information pour proposer des modèles qui perdent leur interprétabilités avec le succès des forêts aléatoires. Les procédures bayésiennes reposent sur les probabilités et inspirent en retour nombre de théories du fonctionnement du cerveau. Elles reposent sur l'idée de causalité et leur modèle est interprétable en ce qu'il repose sur la représentation des influences entre variables. Les algorithmes génétiques s'inspirent de l'évolution et du vivant et reposent sur la simulation de mutations et de sélections des paramètres du modèle. Les machines à vecteurs de support reposent sur l'optimisation mathématique et la multiplication des dimensions qui représentent les données. Enfin, les réseaux de neurones et le "deep learning" s'inspirent du fonctionnement du cerveau pour multiplier les couches nonlinéaires d'une procédure d'apprentissage qui étend son domaine des données dans leurs formes les plus brutes à la prédiction ou la classification.

Cette introduction aux principaux algorithmes du *machine learning* et traditions et choix qu'ils représentent nous permet d'envisager de manière quantitative, dans le chapitre suivant, les communautés qu'ils constituent dans le champ académique.

3

CARTOGRAPHIE DES RECHERCHES SUR ET AVEC L'APPRENTISSAGE ARTIFICIEL

SOMMAIRE

Dans le chapitre 2 on a pu observer comment les algorithmes d'apprentissage étaient formulés en s'inspirant de certains domaines tels que l'évolution biologique ou les neurosciences. Leurs objectifs et leur formalisme les rendent également plus ou moins adaptés à certains domaines d'application. Afin d'analyser comment l'apprentissage artificiel s'inscrit et déborde les champs scientifiques traditionnels, ce chapitre adopte une perspective résolument empirique en étudiant systématiquement la structure et la dynamique des publications scientifiques associées. L'étude des publications scientifiques est un moyen privilégié pour observer comment se structurent les dynamiques d'émergence des domaines scientifiques et, à ce titre, peut nous permettre de saisir comment chaque communauté d'algorithme émerge et interagit avec ses domaines d'applications pour construire,

ou non, la cohérence du domaine de recherche de l'apprentissage artificiel.

Afin de pouvoir réaliser cette analyse, notre étude se base sur des corpus de données extrait de *Web of Science*[1] (WOS). Dans un premier temps, on décrit ces corpus et les caractéristiques principales des communautés d'algorithmes d'apprentissage qu'ils représentent (§3.1). Dans un deuxième temps, on présente la méthode d'analyse des réseaux de co-citations de chaque corpus qui vise à faire émerger et identifier les structures disciplinaires qui les sous-tendent (§3.2). Enfin, on présente chaque thématique produite par ces réseaux afin de pouvoir analyser leur démographie et leur articulation les unes avec les autres à l'aide d'une analyse de la co-présence des auteurs qui les animent (§3.3).

3.1 EXTRACTION ET CARACTÉRISTIQUES PRINCIPALES DES CORPUS

Afin de pouvoir analyser les traces des auteurs de publications scientifiques faisant usage des techniques d'apprentissage, on extrait plusieurs corpus depuis une plateforme de données bibliométrique (§3.1.1). Ces corpus nous permettent de faire des premières analyses descriptives, d'abord sur les auteurs et leurs publications (§3.1.2), puis sur les pays et les domaines d'intérêts les plus représentés (§3.1.3).

3.1.1 Web Of Science *et ses corpus de données*

Web of Science (WOS) est une plateforme développée par Thomson Reuters qui indexe les articles scientifiques, en extrait et structure plusieurs informations telles que : les articles cités, les auteurs, les mots-clés utilisés, le résumé, le titre, les affiliations des auteurs, etc. Alors que l'accès à WOS est payant et l'ensemble de la plateforme fermée au grand public, il existe plusieurs alternatives qui enrichissent leurs bases de données des contributions de leurs utilisateurs. C'est notamment le cas de *Google Scholar* qui, selon une estimation de mai 2014 [58], référence plus de 160 millions d'entrées, couvrant ainsi 80 à 90% de tous les articles publiés en anglais. Néanmoins, ce service ne met à disposition aucun moyen d'extraire ses données (API, exports). Notre étude repose donc sur *Web of Science*. En effet, WOS permet d'extraire plusieurs centaines d'entrées de manière structurée et un script développé par l'équipe Cortext[2], hébergée dans mon laboratoire de thèse (LISIS–INRA), permet d'étendre cette possibilité à des corpus de

1. https://webofknowledge.com/
2. http://cortext.net/

Nom	Requête	Date	Nombre d'entrées
Machine Learning	TS=("machine learning")	Juillet 2016	31,079
Decision Tree	TS=("decision tree*")	Juillet 2016	17,379
Random Forest	TS=("random forest*")	Juillet 2016	5,635
Naive Bayes	TS=("naive bayes*")	Février 2016	4,536
Bayes Net	TS=("bayesian network*")	Février 2016	9,915
Genetic Algorithm	TS=("genetic algorithm*")	Juillet 2016	86,775
SVM	TS=("support vector machine*")	Février 2016	42,498
Neural Net	TS=("artificial neural network*")	Juillet 2016	52,267

Tableau 6 – Résumé des corpus extraits de *Web of Science*.

plusieurs dizaines de milliers de références. Même si le nombre de journaux couverts par le WOS n'est pas exhaustif, la sélection de ceux-ci garantit une bonne couverture de l'ensemble des domaines académiques (à l'exception peut-être des sciences humaines et sociales) dont les revues les plus centrales sont systématiquement indexées. De plus, comme on l'a déjà précisé, les méta-données associées aux articles ont l'avantage d'être extrêmement riches et déjà nettoyées : on peut ainsi aisément identifier les pays de publication d'un article, ou les journaux cités.

Les corpus étudiés ici partent de l'idée simple qu'un moyen pertinent pour regrouper les publications ayant trait à l'apprentissage artificiel et à ses algorithmes, est de chercher les articles qui font mention de leurs dénominations les plus courantes. WOS permet de faire un telle requête générale avec l'option "Topic" (TS) qui cherche l'occurrence de la chaine de caractères passée en paramètre dans l'ensemble du contenu textuel disponible de chaque entrée bibliographique (titre, résumé, mots-clés). De plus, WOS permet d'utiliser un système d'expressions rationnelles afin de capturer des variations de la requête. Par exemple, un requête réalisée avec l'expression "naive bayes*" rassemblera à la fois les articles mentionnant "naive bayes" ou "naive bayesian classifier". Le Tableau 6 rassemble les requêtes à l'origine de chaque corpus extrait de WOS ainsi que le nombre de références qu'elles ont permis d'extraire.

Le corpus *Machine Learning* rassemble l'ensemble des publications faisant référence à ce domaine de recherche sans forcément faire appel à un des algorithmes que nous considérons. Ainsi, il comprend les publications traitant d'autres algorithmes, de problématiques transversales, d'applications, d'implémentations, etc. On peut observer un écart important entre ce corpus général et les corpus *Genetic Algorithm*, *Neural Net* et *SVM* dont les volumes de publications sont largement supérieurs. Si certains algorithmes, comme les algorithmes gé-

nétiques, sont plutôt délaissés aujourd'hui, notre requête qui remonte à 1990[3] permet d'illustrer leur importance passée.

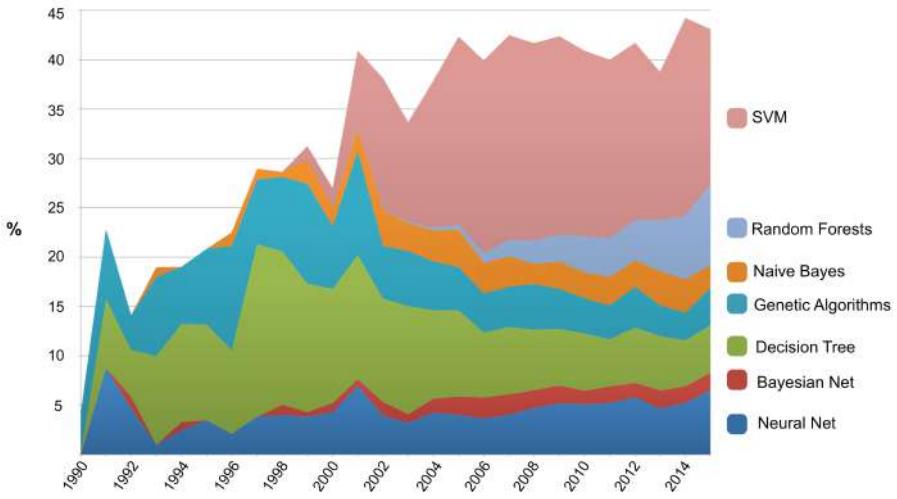


FIGURE 16 – Recouvrement par année de chaque corpus d'algorithme avec le corpus *Machine Learning*.

Afin de circonscrire la place de chaque algorithme sous l'empire de l'apprentissage, la Figure 16 permet d'observer la place de chacun dans le corpus général depuis 1990. L'année 1990 peut probablement être relativisée par des défauts de référencement de WOS, qui sont fréquents jusqu'à cette date. En ignorant les SVM, on voit que l'ensemble des autres algorithmes oscille entre 15% et 30% d'occupation de l'espace du corpus. Rapidement après leur publication à la fin des années 90, ce sont les SVM qui font accroitre significativement la part des algorithmes dans le corpus, cet algorithme représentant à lui seul 30% de nos 7 algorithmes mentionnés en 2006. Une autre évolution remarquable est la croissance importante des forêts aléatoires qui, encore absentes en 2000, occupe plus de 5% de l'espace du corpus général en 2015. Ces fortes croissances ne se font pas au détriment des autres algorithmes dont l'évolution est moins spectaculaire. On peut noter toutefois une légère diminution de la place des algorithmes génétiques et des arbres de décisions qui étaient les principaux algorithmes présents dans les années 1990. Enfin, on peut noter une légère augmentation des réseaux bayésiens et des réseaux de neurones.

3. Les abstracts n'étant systématiquement indexés par le WOS que depuis 1990, il est délicat de prolonger les corpus dans le passé. Nos requêtes portant sur le champ Topic, l'absence d'abstract dans la base diminue « mécaniquement » le nombre de notices identifiables avant cette date.

3.1.2 *Auteurs et publications*

On trouve dans l'ensemble du corpus *Machine Learning* plus de 58,000 auteurs, c'est à dire presque deux fois plus d'auteurs que de publications. Ce chiffre varie sensiblement d'un corpus à un autre, avec 1.2 auteurs par publication en moyenne pour les algorithmes génétiques, 1.4 pour les SVM, 1.8 pour les réseaux de neurones et bayésiens, environ 2.5 pour le naïf bayésien et les arbres de décisions, et 3.1 pour les forêts aléatoires. On explique généralement cet écart par les différences de pratique entre les disciplines de recherche. En effet, le nombre d'auteurs par publication est généralement très élevé en biologie où le travail de groupe en laboratoire est souvent nécessaire [87]. À l'opposé, 66% des publications en mathématiques sont signées par un seul auteur [42]. Cet écart peut être à la fois le fait de contraintes dans les conditions de recherche d'une discipline, par exemple la difficulté d'entreprendre des expériences, mais aussi de la culture qu'une discipline a de promouvoir le travail collectif ou la performance individuelle. Ces tendances générales permettent d'inférer l'écart de nombre d'auteur moyen par publication qui tient probablement à la présence en leurs seins de communautés scientifiques aux pratiques diverses. Cependant, cet indicateur ne nous permet pas d'évaluer le renouvellement des auteurs dans les champs de recherche dont nous souhaitons rendre compte à l'aide des corpus.

Afin de pouvoir mesurer à quel point la forte croissance du nombre de publications est l'œuvre d'un groupe de chercheurs restreint ou d'un flux continu de nouveaux entrants, on peut la comparer avec le pourcentage annuel de nouveaux auteurs qui apparaissent dans le corpus. La Figure 17 prend en compte tous les auteurs du corpus *machine learning* et permet d'observer que le taux de nouveaux entrants décroît progressivement à mesure que le nombre de publications augmente. Cette courbe atteint en 2015 un ratio d'un auteur sur deux n'ayant jamais été observé auparavant. Cela laisse imaginer une structuration progressive de cette thématique de recherche autour d'un nombre important de chercheurs et un nombre croissant de chercheurs plus éphémères, tel que des doctorants et post-doctorants, que le domaine attire mais qui ne restent pas tous dans leur branche académique.

La Figure 18 répète la mesure de ce ratio annuel pour chaque corpus d'algorithme et montre qu'il varie sensiblement de l'un à l'autre. Ainsi, en 2015, la part des nouveaux auteurs varie entre 30% pour les algorithmes génétiques, à près de 70% pour le naïf bayésien et les forêts aléatoires. Une explication possible de cet écart est qu'en parallèle de la structuration du champ de recherche propre à chaque algorithme que les courbes dévoilent, le succès de certains au-delà de leurs frontières accroît le nombre d'utilisateurs intéressés par un

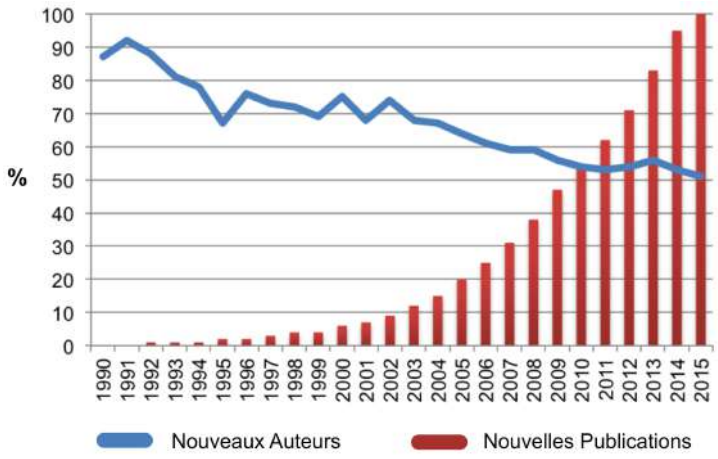


FIGURE 17 – Statistiques de la population globale dans le temps de la communauté du *Machine Learning* : *ratio annuel cumulé de nouvelles publications (histogramme rouge) et ratio annuel de nouveaux auteurs (ligne bleue).*

usage précis et souvent unique plutôt que par l'algorithme lui-même. Ainsi, un ratio élevé de nouveaux auteurs serait un témoin du succès d'un algorithme remobilisé par des chercheurs extérieurs au domaine qui en font un usage purement applicatif. En ce sens, on retrouve parmi les algorithmes ayant ce ratio élevé ceux qui ne nécessitent que peu de paramètres de configuration (arbres de décision, forêts aléatoire, naïf bayésien). À l'opposé on trouve des algorithmes plus complexes, dont il est difficile de produire un classifieur générique, comme les réseaux de neurones ou les algorithmes génétiques.

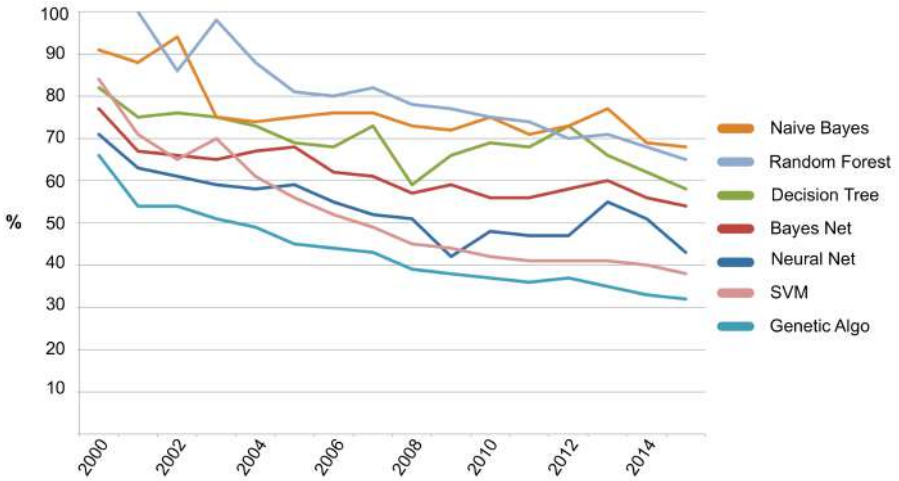


FIGURE 18 – Ratio annuel de nouveaux auteurs dans chaque corpus.

La Figure 17 montre aussi à quel point la recherche récente faisant référence au *machine learning* constitue une majorité dans le corpus. Ainsi, les articles publiés après l'an 2000 constituent 94% du corpus, et ceux publiés après 2010, presque la moitié du corpus. Si ce rythme

de publications se maintenait à cette allure exponentielle, on peut imaginer que la recherche des quelques années à venir occuperait plus de place dans les publications de ce champ de recherche que tout ce qui a été fait jusque là. Cependant cette observation n'est pas propre à l'apprentissage artificiel et montre seulement que ce domaine est en croissance au sein d'une activité de publication scientifique en forte croissance.

3.1.3 *Pays et domaines d'intérêt*

L'affiliation des auteurs permet de situer les principaux pays actifs dans la recherche sur l'apprentissage artificiel. Pour le seul corpus *Machine Learning*, on trouve sans surprise l'Amérique du nord en tête (33%) qui domine, de manière comparable, la plupart des champs de publications académiques. Suivent les principaux pays européens (21%) et asiatique (18%), avec la Chine qui totalise à elle seule 11%. Environ un tiers des publications sont produites par des auteurs affiliés à une longue traine de pays faiblement représentés ($\leqslant$ 1%). On peut observer une situation similaire pour tous les corpus avec la Chine et les EUA en situation dominante et une présence significative du Royaume-Uni. Les autres pays ne semblent être présents de manière notable que sur un ou deux algorithmes, par exemple la France sur les arbres de décisions, l'Inde et l'Iran sur les réseaux de neurones. Il s'agit donc, dans l'ensemble, d'une pratique largement mondialisée très similaire aux domaines scientifiques importants et actifs que l'on peut observer par ailleurs sur WOS. Mais il s'en dégage des particularités sur certains algorithmes et affiliations qui permettent de saisir des spécificités nationales.

Ces spécificités nationales invitent à observer non plus l'activité des pays par algorithmes mais la répartition des algorithmes dans l'activité de chaque pays. Pour ce faire, on retient donc pour chaque algorithme le pourcentage d'activité de chaque pays via l'affiliation des auteurs, pour retenir dans un deuxième temps le pourcentage de cette part d'activité dans l'activité totale du pays pour tous les corpus. La Figure 19 permet d'observer les résultats de cette méthode et donc d'analyser l'activité de chaque pays sans qu'elle soit écrasée par le poids des pays dominants. Une première impression laissée par cette figure est qu'il y a des spécificités géographiques quant aux algorithmes les plus utilisés. On trouve par exemple plus de 25% de la recherche prise en compte pour la Chine et l'Allemagne, consacrée respectivement aux SVM et aux forêts aléatoires. De la même manière, on retrouve des communautés nationales peu intéressées ($\leqslant$ 10%) par certains algorithmes, comme la France par le naïf bayésien, l'Allemagne par les algorithmes génétiques, le Japon par les forêts aléatoires. Enfin, deux algorithmes semblent occuper une place

importante dans la plupart des pays considérés, les réseaux bayésiens et les arbres de décisions.

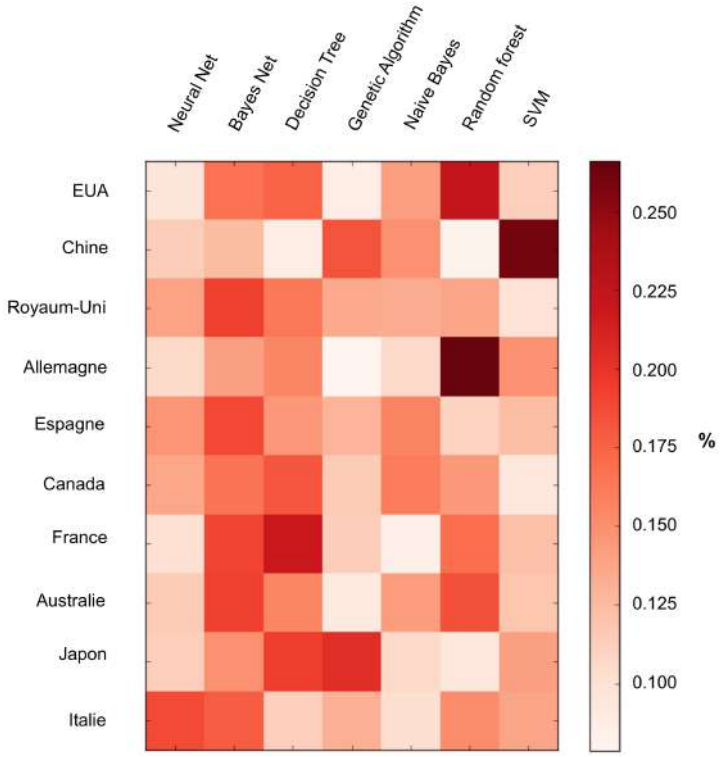


FIGURE 19 – Présence de chaque corpus par pays.

La classification disciplinaire que propose WOS des journaux dans lesquels sont publiés les articles considérés, permet d'observer quels "domaines d'intérêts" (*subject areas*) sont les plus concernés par chaque corpus. On trouve principalement, dans le corpus *Machine Learning*, l'informatique (37%), l'ingénierie (16%), puis une suite de sciences assez variée, comme Biochimie et biologie moléculaire (4%), Mathématiques (3%), Systèmes de contrôle (3%), Imagerie (2%). La prédominance de l'ingénierie de l'informatique se vérifie dans le classement des publications pour tous les autres algorithmes.

Avec la même méthode que pour les pays, la Figure 20 permet d'observer la place de chaque algorithmes pour les 10 domaines les plus fréquents. On peut voir que les SVM et les réseaux bayésiens sont les deux seuls algorithmes à être utilisés de manière constante d'un domaine à un autre. À l'inverse, les forêts aléatoires et les algorithmes génétiques semblent montrer une variabilité bien plus importante d'usage, avec notamment une forte présence des forêts aléatoires pour tous les domaines en lien avec la biologie, qui font par ailleurs peu usage des algorithmes génétiques. De manière générale, à part les deux domaines principaux que sont l'informatique et l'ingénierie, chaque domaine montre une empreinte particulière de l'utilisation des divers algorithmes représentés par nos corpus.

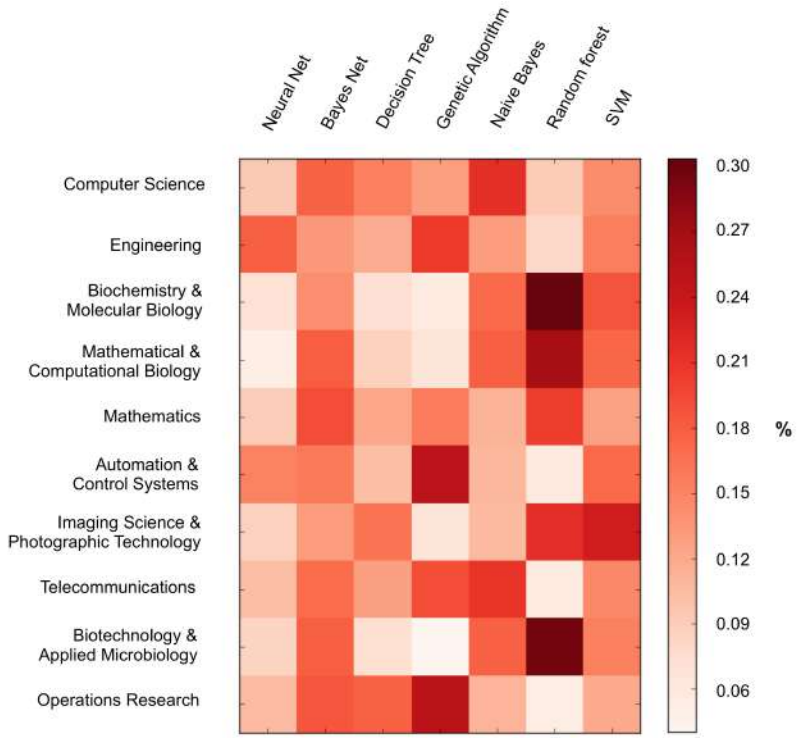


FIGURE 20 – Présence de chaque corpus par domaine d'intérêt .

Néanmoins, les catégories sur lesquelles reposent WOS, comme la plupart des outils traditionnels de bibliométrie, sont définies *a priori*, c'est à dire qu'elles s'appuient sur des systèmes consensuels et relativement figés de classification de l'activité scientifique essentiellement guidés par les besoins en recherche d'information hérités de la recherche documentaire. Ainsi, elles peuvent se montrer peu efficaces pour observer les processus progressifs de structuration et l'hétérogénéité des domaines de recherches émergents comme celui dont nous essayons de rendre compte ici. Comme nous avons pu le voir dans le chapitre précédent, le *machine learning* n'est pas une discipline en tant que telle, mais un croisement entre des disciplines existantes, et si quelques mots-clés peuvent récemment venir l'indexer dans des systèmes bibliographiques, ce ne serait que sur des corpus très récents. Les statistiques descriptives basées sur le système de classification de WOS que nous avons présentées dans cette section écrasent la diversité de l'apprentissage sous le poids de ses affiliations classiques (informatique, ingénierie) et peinent à révéler ses spécificités. Elles échouent à montrer comment des sous-ensembles internes à ce champ structurent celui-ci et s'appuient sur des catégories qui ne sont pas pertinentes pour saisir des mouvements récents et les origines thématiques des auteurs. Pour cela, nous faisons appel à un ensemble d'outils plus adaptés pour remplir cet objectif qui nous permette de saisir comment chaque publication se projette elle-même sur

un champ de recherche dont l'union constitue un réseau de relation dans lequel émergent les thématiques.

3.2 MÉTHODOLOGIE DE RECONSTRUCTION DES THÉMATIQUES DE L'APPRENTISSAGE ARTIFICIEL

Afin de définir une méthode de reconstruction des domaines thématiques de l'apprentissage, on envisage l'activité des chercheurs en partant des références qu'ils citent dans leurs publications et qui, en elles-mêmes, fournissent de nombreuses informations sur les différentes communautés algorithmiques envisagées (§3.2.1). On décrit ensuite comment on construit les réseaux de co-citations et la méthode par laquelle on identifie les thématiques dominantes de chaque corpus (§3.2.2).

3.2.1 *Citations*

En prenant un peu de recul par rapport aux données disponibles dans les corpus WOS et aux statistiques descriptives qu'elles nous ont permis de produire dans la section précédente, on peut envisager une publication comme un ancrage dans une ou plusieurs communautés de recherche. Ainsi, un modèle simple de l'activité scientifique est de considérer chaque publication comme une contribution à un état de l'art qui caractérise ce à quoi on s'identifie en tant qu'auteur. Lorsque un article cite d'autres articles, il décrit son ancrage formel dans un champ de connaissance qui définit le domaine de sa contribution. Cet environnement porte une dimension subjective en ce qu'il représente comment le ou les auteurs envisagent leurs contextes et frontières intellectuels et la manière dont ils peuvent préciser leurs ambitions. Cet ancrage formel porte aussi une dimension objective en ce que chacune de ces références a une place définie dans les différentes structures qui composent le monde de la recherche comme les journaux, les disciplines, les départements, les financements, etc.

Afin d'illustrer ce propos on peut simplement lister des échantillons de références de tous les corpus selon différents critères de sélection. La Tableau 7 liste les 5 publications les plus citées dans chaque corpus et fait ressortir ainsi la plupart des travaux mentionnés dans le chapitre précédent pour retracer l'historique de chaque algorithme considéré. On retrouve alors Quinlan et Breiman pour les arbres de décision et les forêts aléatoires, Pearl et Friedman pour les approches bayésiennes, Rumelhart pour les réseaux de neurones, Goldberg et Holland pour les algorithmes génétiques, et Vapnik pour les SVM. Pour le corpus *Machine Learning*, les auteurs les plus cités semblent

illustrer les communautés qui croissent le plus vite en son sein, c'est à dire principalement les forêts aléatoires et les SVM, comme illustré précédemment par la Figure 16.

La liste complète des auteurs cités dans le chapitre précédent met en lumière les travaux complémentaires souvent sollicités pour appuyer un propos. Ainsi on retrouve certaines publications ayant trait à des champs d'application caractéristique d'un ou plusieurs algorithmes comme l'écologie pour les forêts aléatoires (Cutler, 2007), la reconnaissance de motifs pour les réseaux de neurones, les naifs bayésiens et les SVM (Duda, 1973 ; Bishop, 1995 ; Burges, 1998). Aussi, on retrouve des publications traitant des contraintes d'apprentissage que l'algorithme considéré est connu pour savoir habilement gérer, par exemple l'apprentissage multi-objectif pour les algorithmes génétiques (Deb, 2002).

Si les références les plus citées nous permettent de mettre en lumière les recherches ayant formulées les hypothèses mentionnées dans chaque corpus, on peut aussi attendre des références citées qu'elles nous informent sur les traditions scientifiques sur lesquelles celles-ci reposent. Dans ce sens, le Tableau 8 liste les 5 références les plus anciennes parmi les 1000 plus citées, permettant ainsi de mettre en valeur les travaux de recherche anciens qui sont fréquemment sollicités dans la formulation de la recherche contemporaine. Les listes propres à chaque corpus se recouvrent largement. Par exemple, la publication de Fischer (1936) qui introduit la base de donnée *Iris*, présentée dans le premier chapitre de cette thèse (§1.1.1.2), est omniprésente et se retrouve dans presque tous les corpus. Le fait qu'une base de données soit la référence la plus partagée entre les différents corpus montre combien la question de la classification est centrale pour l'ensemble de ces algorithmes. Cette base de données est en effet très utilisée pour démontrer les propriétés d'un algorithme et la capacité ou limite de celui-ci à séparer correctement les classes d'espèces déjà connues. Elle montre également le rapport naturel des différents algorithmes d'apprentissage à l'empirisme qui passe par des méthodes d'évaluation les rendant comparables les unes aux autres. Ensuite, on trouve des références à des travaux de statistiques fondateurs qui plongent dans des clivages anciens, par exemple, Pearson pour les approches fréquentistes (SVM, réseaux de neurones) et Bayes pour les approches éponymes (réseau et naïf). Bien qu'absent de ce clivage, le travail de Darwin apparaît comme fondateur pour les algorithmes génétiques. Enfin on voit que la théorie de l'information de Shanon a un ancrage dans plusieurs de ces communautés, notamment les arbres de décision et les forêts aléatoires. On voit donc se recouper, dans cet ensemble de corpus, des références à des recherches séminales de différentes branches des sciences et techniques ce qui nous invite à considérer la communauté qui nous intéresse ici, l'apprentis-

Corpus	1er Auteur	Date	Titre
Machine Learning	Quinlan	1993	*C4.5 : Programs for machine learning*
	Vapnik	1998	*Statistical learning theory*
	Vapnik	1995	*The nature of statistical learning theory*
	Breiman	2001	*Random forests*
	Mitchell	1997	*Machine learning*
Decision Tree	Quinlan	1993	*C4.5 : Programs for machine learning*
	Breiman	1984	*Classification and regression trees*
	Quinlan	1986	*Induction of decision trees*
	Breiman	1996	*Bagging predictors*
	Witten	2005	*Data mining : pratical machine learning tools and techniques*
Random Forest	Breiman	2001	*Random forests*
	Liaw	2002	*Classification and regression by randomForest*
	Breiman	1984	*Classification and regression trees*
	Breiman	1996	*Bagging predictors*
	Cutler	2007	*Random forests for classification in ecology*
Bayesian Net	Pearl	1988	*Probabilistic reasoning in intelligent systems*
	Cooper	1992	*A Bayesian method for the induction of probabilistic networks from data*
	Heckerman	1995	*Learning Bayesian networks : The combination of knowledge and statistical data*
	Friedman	1997	*Bayesian network classifiers*
	Jensen	1996	*An introduction to Bayesian networks*
Naive Bayes	Friedman	1997	*Bayesian network classifiers*
	Domingos	1997	*On the optimality of the simple Bayesian classifier under zero-one loss*
	Mitchell	1997	*Machine learning*
	Witten	2005	*Data mining : pratical machine learning tools and techniques*
	Pearl	1988	*Probabilistic reasoning in intelligent systems*
Neural Network	Rumelhart	1986	*Parallel Distributed Processing*
	Haykin	1999	*Neural Networks*
	Bishop	1995	*Neural Networks for Pattern Recognition*
	Hornik	1989	*Multilayer feedforward networks*
	Rumelhart	1986	*Learning Internal Representations*
Genetic Algorithm	Goldberg	1989	*Genetic Algorithms in search, optimization and machine learning*
	Holland	1975	*Adaptation in natural and artificial systems*
	Deb	2002	*A fast and elitist multiobjective genetic algorithm*
	Davis	1991	*Handbook of genetic algorithms*
	Kirkpatrick	1983	*Optimization by simulated annealing*
SVM	Vapnik	1995	*The nature of statistical learning theory*
	Vapnik	1998	*Statistical learning theory*
	Cortes	1995	*Support vector networks*
	Burges	1998	*A tutorial on SVM for pattern recognition*
	Cristianini	2000	*An introduction to SVM*

Tableau 7 – Les 5 références les plus citées dans chaque corpus.

sage artificiel, comme une coupe transversale en leurs seins, avec des objectifs et problématiques propres à son champ de recherche que les références les plus récentes peuvent illustrer.

Afin de saisir les éléments les plus caractéristiques de la recherche contemporaine dans chaque corpus, la Tableau 9 liste les 5 références les plus récentes parmi les 1000 les plus citées. On voit d'abord que les dates qui ressortent de cette sélection varient d'un corpus à l'autre. Les algorithmes les plus anciens font émerger des publications de 2010 et 2011 (algorithmes génétiques, réseaux de neurones), alors que ceux formulés plus récemment pointent vers des publications datant de 2013, 2012 (random forests, SVM). On trouve beaucoup moins de publications transversales que pour le Tableau 8 et les travaux cités dans plusieurs corpus concernent soit des bases de données (Bache, 2013), un langage de programmation (R Team, 2013) ou un livre d'enseignement du "data mining" (Han, 2012). Ainsi, on peut observer que l'unité thématique des différents corpus se construit surtout autour des problématiques d'enseignement et d'implémentation mais que le reste de la recherche sur un algorithme semble rester spécifique à ses contraintes. Les publications singulières à chaque corpus traitent de questions diverses telle que l'exploration de nouveaux champs d'application (Dias, 2013 ; Wouter, 2013), de variations des algorithmes concernés pour certaines contextes scientifiques et un nombre important de revues de l'état de l'art.

Ce paysage des publications récentes ou anciennes, rendu possible par ces listes simples, propose l'image d'un champ actif, en train de se structurer autour de certaines ressources (données, enseignement, outils) et d'explorer une variété importante de contraintes plus sensibles à la communauté d'algorithme concernée. Afin de poursuivre cette observation avec une granularité plus fine d'analyse, et surtout de ne plus dépendre des catégories de WOS pour caractériser les domaines scientifiques concernés, la section suivante fait appel aux analyses de réseaux.

3.2.2 *Méthodologie d'analyse*

Cette étude scientométrique se concentre sur le processus de structuration des communautés envisagées pour mettre en lumière la composition largement hétérogène du domaine de l'apprentissage artificiel. Les techniques de modélisation et de visualisation utilisées permettent de réaliser la cartographie de leurs citations. Dans un premier temps, nous décrivons un résumé des techniques choisies pour définir les stratégies d'analyse et de cartographie adoptées (§3.2.2.1). Ensuite nous montrons comment, en analysant ce type de réseaux, nous

Corpus	1er Auteur	Date	Titre
Machine Learning	Pearson	1901	*Principal Component Analysis*
	Mercer	1909	*Functions of positive and negative type, and their connection with the theory of integral equations*
	Hotelling	1933	*Analysis of a complex of statistical variables into principal components*
	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	Friedman	1937	*The use of ranks to avoid the assumption of normality implicit in the analysis of variance*
Decision Tree	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	Friedman	1937	*The use of ranks to avoid the assumption of normality implicit in the analysis of variance*
	Von Neumann	1944	*The Theory of Games and Economic Behavior*
	Wilcoxon	1945	*Individual comparaisons by ranking methods*
	Shanon	1948	*A Mathematical Theory of Communication*
Random Forest	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	Friedman	1937	*The use of ranks to avoid the assumption of normality implicit in the analysis of variance*
	Jenny	1941	*Factors of soil formation : a system of quantitative pedology*
	Dice	1945	*Measures of the amount of ecologic association between species*
	Shanon	1948	*A Mathematical Theory of Communication*
Bayesian Net	Bayes	1763	*An essay towards solving a problem in the doctrine of chances*
	Wright	1921	*Correlation and causation*
	Shanon	1948	*A Mathematical Theory of Communication*
	Kullback	1951	*On information and sufficiency*
	Metropolis	1953	*Equation of state calculations by fast computing machines*
Naive Bayes	Bayes	1763	*An essay towards solving a problem in the doctrine of chances*
	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	Friedman	1937	*The use of ranks to avoid the assumption of normality implicit in the analysis of variance*
	Shanon	1948	*A Mathematical Theory of Communication*
	Kullback	1951	*On information and sufficiency*
Neural Network	Pearson	1901	*Principal Component Analysis*
	Hotelling	1933	*Analysis of a complex of statistical variables into principal components*
	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	McCulloch	1943	*A logical calculus of the ideas immanent in nervous activity*
	Levenberg	1944	*A method for the solution of certain non–linear problems in least squares*
Genetic Algorithm	Darwin	1859	*On the origin of the species by natural selection*
	Pareto	1896	*Cours d'économie politique*
	Fisher	1936	*The use of multiple measurements in taxonomic problems*
	Ziegler	1942	*Optimum settings for automatic controllers*
	McCulloch	1943	*A logical calculus of the ideas immanent in nervous activity*
SVM	Pearson	1901	*Principal Component Analysis*
	Mercer	1909	*Functions of positive and negative type, and their connection with the theory of integral equations*
	Hotelling	1933	*Analysis of a complex of statistical variables into principal components*
	Wilcoxon	1945	*Individual comparisons by ranking methods*
	Shanon	1948	*TA Mathematical Theory of Communication*

Tableau 8 – Les 5 plus anciennes références parmi les 1000 plus citées par corpus.

Corpus	1er Auteur	Date	Titre
Machine Learning	Bache	2013	*UCI machine learning repository*
	R Team	2013	*R : A language and environment for statistical computing*
	Huang	2012	*Extreme learning machine for regression and multiclass classification*
	Gaulton	2012	*ChEMBL : a large-scale bioactivity database for drug discovery*
	Orru	2012	*Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease*
Decision Tree	Han	2012	*Data mining : concepts and techniques*
	Witten	2011	*Data mining : pratical machine learning tools and techniques*
	Archarya	2011	*Automatic detection of epileptic EEG signals using higher order cumulant features*
	R Team	2013	*R : A language and environment for statistical computing*
	Bache	2013	*UCI machine learning repository*
Random Forest	R Team	2013	*R : A language and environment for statistical computing*
	Wouter	2013	*Data mining in the Life Sciences with Random Forest : a walk in the park*
	Rodriguez-Galiano	2012	*An assessment of the effectiveness of a random forest classifier for land-cover classification*
	Li	2912	*Prediction of protein domain with mRMR feature selection and analysis*
	Criminisi	2012	*Decision forests : A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning*
Bayesian Net	Dias	2013	*Evidence synthesis for decision making*
	Khakzad	2013	*Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network*
	Pitchforth	2013	*A proposed validation framework for expert elicited Bayesian Networks*
	Khakzad	2013	*Risk-based design of process systems using discrete-time Bayesian networks*
	Nagarajan	2013	*Bayesian networks in R*
Naive Bayes	Bache	2013	*UCI machine learning repository*
	Koutsoukas	2013	*In silico target predictions : defining a benchmarking data set and comparison of performance of the multiclass naïve bayes and parzen-rosenblatt window*
	Zaidi	2013	*Alleviating naive Bayes attribute independence assumption by attribute weighting*
	Han	2012	*Data mining : concepts and techniques*
	Jiang	2012	*Improving Tree augmented Naive Bayes for class probability estimation*
Neural Network	Nazari	2011	*Modeling ductile to brittle transition temperature of functionally graded steels by artificial neural networks*
	Gharagheizi	2011	*Determination of Parachor of Various Compounds Using an Artificial Neural Network*
	Witten	2011	*Data Mining : Practical Machine Learning Tools and Techniques*
	Sezer	2011	*Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping : Klang valley, Malaysia*
	Yilmaz	2010	*Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey*
Genetic Algorithm	Das	2011	*Differential evolution : a survey of the state-of-the-art*
	Zhou	2011	*Multiobjective evolutionary algorithms : A survey of the state-of-the-art*
	Derrac	2011	*A practical tutorial on the use of nonparametric statistical tests as a methodologyfor comparing evolutionary and swarm intelligence algorithms*
	Karaboga	2011	*A novel clustering approach : Artificial Bee Colony (ABC) algorithm*
	Bhattacharya	2010	*Hybrid differential evolution with biogeography-based optimization for solution of economic load dispatch*
SVM	Chen	2013	*iRSpot-PseDNC : identify recombination spots with pseudo dinucleotide composition*
	Bache	2013	*UCI machine learning repository*
	Xu	2013	*Analysis of a complex of statistical variables into principal components*
	Huang	2012	*iSNO-PseAAC : Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition*
	Chou	2012	*iLoc-Hum : using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites*

Tableau 9 – Les 5 références les plus récentes parmi les 1000 plus citées par corpus.

parvenons à dégager les communautés thématiques sous-jacente propres à chaque algorithme (§3.2.2.2).

3.2.2.1 *Méthode de construction des réseaux de co-citations*

Comme nous l'avons vu, choisir de représenter les citations d'une publication plutôt que des informations sur la publication elle-même, comme ses auteurs ou des mots-clés, permet de faire émerger son ancrage formel dans un champ de connaissance. C'est pour ces raisons que l'analyse des co-citations est notamment une méthode reconnue d'analyse en scientométrie [123]. Par co-citation on entend la présence de deux éléments cités dans le même article. Comme élément, ou *nœud* du réseau, nous choisissons ici le journal cité, ce qui permet d'abstraire un peu les thématiques des articles cités en se basant sur les classements et processus de sélection de chaque journal. Cette abstraction offre aussi plus de visibilité lors de l'observation du réseau, où le titre d'un journal peut être plus évocateur que les titres isolés des nombreux papiers qu'il publie. La taille de chaque nœud est proportionnelle à la fréquence de citation du journal. Lorsque deux journaux sont cités conjointement par un nombre important d'articles, un lien est créé entre eux dans le réseau (un « nombre important » étant naturellement à relativiser par la fréquence respective de citation de chaque journal).

Cette méthode offre alors la possibilité de visualiser des groupes de citations fréquemment co-cités. Afin de déterminer ces groupes, les réseaux sont analysés par un algorithme de détection de communautés, aussi appelé algorithme de regroupement (*clustering*). Plus spécifiquement, l'analyse présentée ici produit cette structure haut-niveau du graphe grâce à l'algorithme de Louvain [14]. Chaque cluster fait l'objet d'une coloration propre dont la taille du cercle sous-jacent est proportionnelle au nombre d'articles dans le corpus qui se projettent sur ce groupe de journaux cités. La procédure de projection utilisée retient au plus un cluster d'appartenance par article. C'est donc à une seule « communauté épistémique » de prédilection que sont assignés les articles. Si le recouvrement entre les journaux cités d'un article et la composition des clusters est ambigüe, aucune communauté d'appartenance n'est assignée à l'article. Enfin, un algorithme de spatialisation permet de disposer les nœuds de manière à ce que leur voisinage géométrique soit le plus proche possible de leur voisinage topologique.

L'ensemble des méthodes décrites dans cette section ont été implémentés avec les outils regroupés sur la plateforme Cortext[4], développée dans mon laboratoire de thèse (LISIS–INRA). Plusieurs publi-

4. http ://cortext.net

cations détaillent déjà ces méthodes avec de plus amples détails [104, 113].

3.2.2.2 *Méthode d'identification des communautés thématiques*

Afin d'explorer un premier exemple minimal, la Figure 21 permet la visualisation du réseau des co-citations de journaux sur le corpus *Machine Learning*. A des fins d'illustration de la méthode, celle-ci est volontairement limitée à un nombre réduit de nœuds (30), afin de garantir une certaine lisibilité des labels.

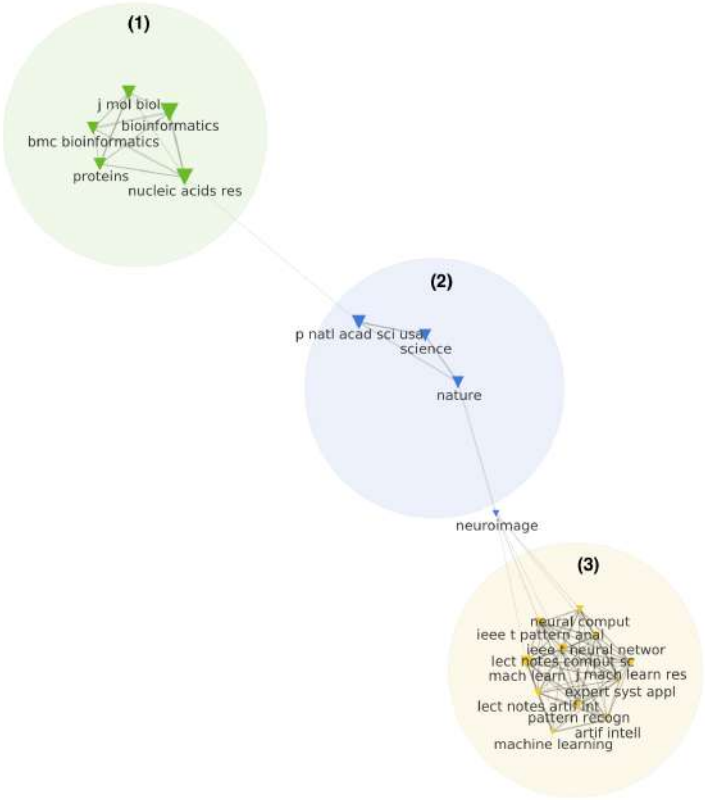


FIGURE 21 – Réseau de co-citations des 30 journaux les plus cités dans le corpus *Machine Learning*.

L'algorithme de *clustering* met très distinctement trois groupes en lumière. Une première méthode pour identifier disciplinairement ces groupes est d'observer les journaux cités en leur sein. Ainsi, le *groupe 1* ne comporte que des journaux de biologie, notamment de bioinformatique (*Bioinformatics*), biologie moléculaire (*Journal of molecular biology*). On imagine assez bien que les travaux représentés par ce groupe sont surtout des utilisations de techniques d'apprentissage, par exemple, pour la prédiction de structure des protéines, ou l'analyse des réseaux de gènes. Le *groupe 3* représente le cœur de métier de la recherche en apprentissage artificiel. Son objet premier est l'amé-

lioration ou l'invention de techniques de prédiction ou de classification. Comme nous le verrons, tous les réseaux de nos corpus d'algorithmes contiennent au moins un groupe qui représente le cœur de la recherche sur l'algorithme lui-même. Ici, on retrouve de manière assez attendue au centre du groupe le journal *Machine Learning* et plusieurs autres qui font référence à des sous-ensembles de ce domaine, soit en terme applicatif (IEEE *transactions in pattern analysis*), soit en terme spécifique en désignant certains algorithmes (IEEE *transactions in neural networks*). Enfin, on trouve des revues qui englobent la catégorie de l'apprentissage, comme *Artificial intelligence*. Ce domaine se place donc en amont d'applications concrètes à des données de terrain qui apparaissent ici surtout comme un support pour démontrer l'efficacité ou l'originalité d'un algorithme. Le *groupe 2* réunit des journaux pluridisciplinaires à fort impact comme *Science*, *Nature* ou PNAS qui font le pont entre les deux autres groupes qui ne partagent, par ailleurs, aucun lien.

En limitant notre observation à ce réseau, on est tenté d'émettre plusieurs hypothèses à partir des communautés qui ont été détectées, et de leurs topologies. Ainsi, la présence d'une seule communauté thématique d'application, la biologie, semble placer cette discipline comme principal champ d'utilisation des techniques d'apprentissage. La topologie du graphe, ie. le placement des communautés et de leurs connections, permet d'observer que la biologie interagit très peu avec la recherche fondamentale en apprentissage, si ce n'est pour partager leurs succès dans les mêmes journaux interdisciplinaires à forts impacts, représentés par le groupe 2. On peut donc imaginer une relation unilatérale, où une partie de la recherche en biologie utilise les techniques de *machine learning* sans vraiment collaborer avec les chercheurs du domaine dont ces techniques sont issues. Cependant, comme nous l'avons vu dans les chapitres précédents, la relation entre ces deux champs de recherche n'est pas si caricaturale, et l'apprentissage artificiel reprend plusieurs inspirations du monde du vivant pour explorer des pistes d'imitation et de simulation du vivant, de l'adaptation à des contraintes, de l'intelligence. Ainsi, les publications de biologie qui apparaissent dans les revues à fort impact peuvent aussi, en retour, inspirer des pistes de recherche pour l'apprentissage.

Ce premier réseau est volontairement de taille limitée afin d'augmenter sa lisibilité et d'illustrer de manière schématique la méthode décrite précédemment. Cependant sa taille réduite comporte de nombreux biais, notamment celui d'être beaucoup plus sensible à la surreprésentation de disciplines par ailleurs très actives dans le monde académique, comme la biologie. Pour permettre à davantage de communautés d'émerger et afin de pouvoir raffiner les hypothèses concer-

nant leur nature et leur place dans le réseau, il est nécessaire d'augmenter la taille de celui-ci, quitte à perdre en lisibilité.

Dans ce sens, la Figure 22 permet d'observer le réseau de co-citation des 200 journaux les plus cités dans le corpus *Machine Learning* et donc de voir émerger plus de communautés et une topologie plus nuancée pour caractériser la recherche faisant mention de l'apprentissage artificiel. Lorsque la densité des liens au sein d'une communauté est trop élevée, il est difficile d'accéder via la visualisation à la lecture des titres de journaux. Dans ce cas, on y accède en aparté, en réalisant directement cette requête sur les données ayant produit le graphe.

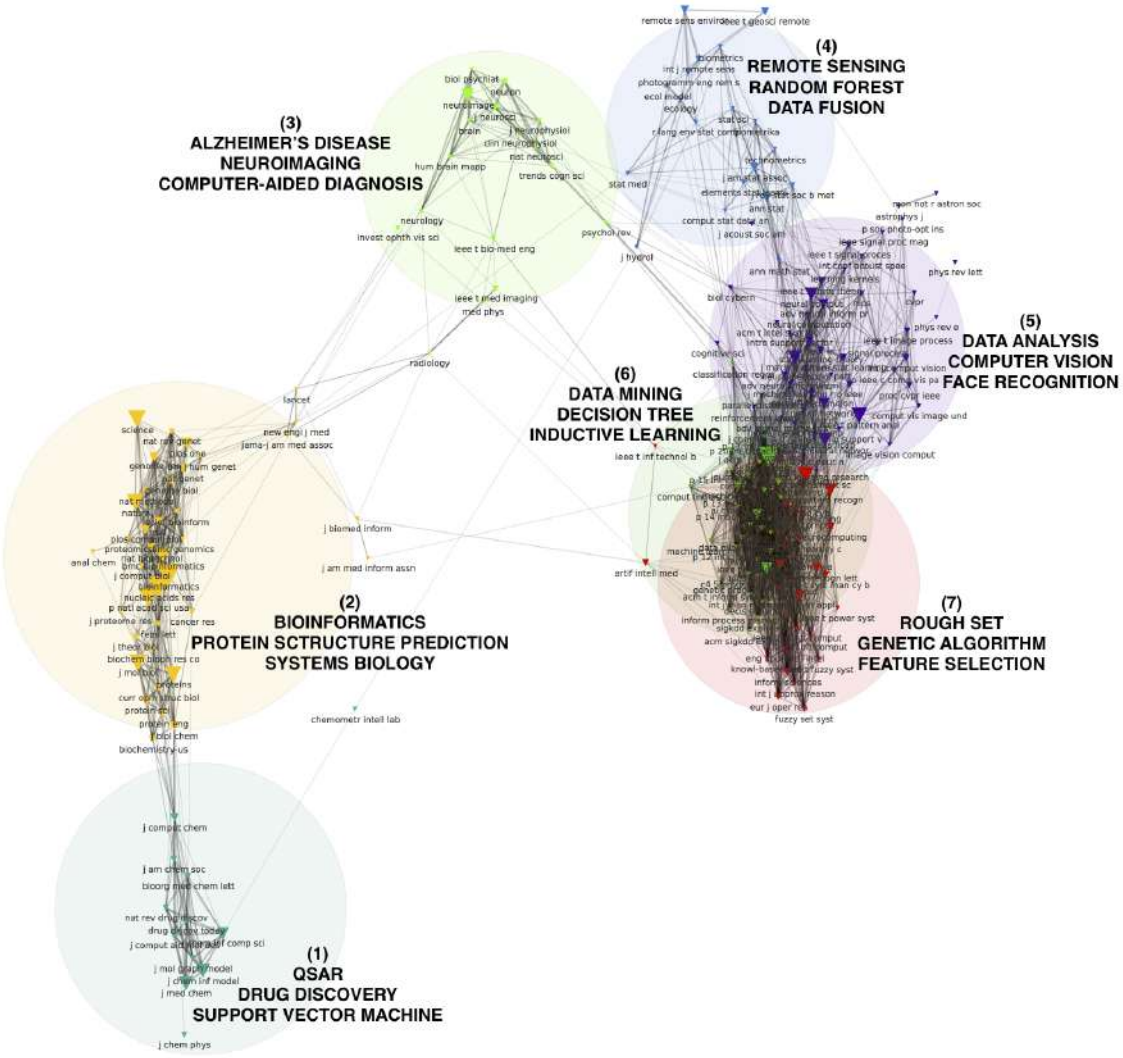


FIGURE 22 – Réseau de co-citations des 200 journaux les plus cités dans le corpus *Machine Learning*.

Chaque communauté est assortie des 3 mots-clés les plus fréquents en son sein, ce qui permet, en complément de la lecture des journaux les plus centraux, de caractériser qualitativement le domaine de

recherche concerné. On peut déceler des éléments d'information supplémentaires concernant un groupe comme des domaines fréquents d'application, des termes d'identification des techniques d'apprentissage caractéristiques d'une discipline. Par exemple, dans la Figure 22, le groupe 3 est assez facilement identifiable aux neurosciences, notamment par la position centrale de journaux comme *Neuroimage*, *Journal of neuroscience*, etc. Les mots-clés font ressortir un cas d'utilisation représentatif de l'apprentissage en neurosciences qu'est le diagnostic assisté par ordinateur (*comptuter-aided diagnosis*) de la maladie d'Alzheimer (*alzheimer's disease*) via la détection d'éléments pertinents sur des images notamment issues d'IRM ou d'EEG (*neuroimaging*). De la même manière, Le groupe 2, qui concerne la biologie, fait ressortir le cas d'utilisation de l'apprentissage pour la prédiction de la structure des protéines et des sous-domaines thématiques comme la bioinformatique et la biologie des systèmes.

Dans certains cas, le fait de disposer à la fois des mots-clés et des titres de journaux permet aussi de préciser la nature d'une communauté dont la présence d'un seul de ces éléments laisserait persister des ambiguïtés. Par exemple, dans le cas du groupe 5, les titres des journaux cités sont assez hétérogènes. Alors que certains font référence à l'analyse d'image comme IEEE *transactions on image processing*, on trouve aussi plusieurs sous-communautés comme l'astrophysique (*Astrophysics journal*), la physique (*Physical review letters*), etc. Dans ce cas, les mots clés *computer vision* ou *face recognition* permettent de classer avec moins d'ambiguïté ces sous-communautés relevant du traitement de l'image et de labelliser ce groupe sous cette thématique. Inversement, les mots-clés des groupes 6 et 7 ne permettent pas de déterminer de quelle communauté thématique il s'agit. Observer les journaux les plus centraux dans chacune (*machine learning*, *machine learning research*, *lectures notes in computer sciences*) permet de qualifier ces communautés comme relevant de recherches fondamentales sur les algorithmes et non comme un domaine d'application particulier.

3.3 LES DOMAINES DE RECHERCHE ET D'APPLICATIONS DE L'APPRENTISSAGE

Dans la section précédente, nous avons exposé un ensemble de méthodes pour représenter les réseaux de co-citations des corpus WOS et qualifier thématiquement les communautés qu'elles révélent. Dans cette section, on applique cette méthode sur tous les corpus afin de pouvoir décrire les principales communautés que l'on retrouve pour tous les algorithmes considérés (§3.3.1). Cette identification systématique nous permet d'observer comment chaque communauté d'algorithme se structure dans le temps (§3.3.3). De plus, cette typologie commune des thématiques du *machine learning* nous permet de

confondre les corpus dans l'analyse et donc de pouvoir juger si ce sont les thématiques ou les algorithmes qui déterminent le plus le déplacement des auteurs (§3.3.2).

3.3.1 *Les thématiques de chaque algorithme*

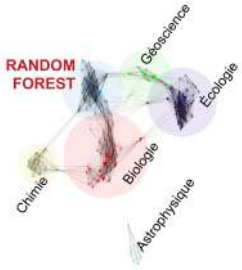
C'est en suivant la méthode d'identification des communautés que nous avons décrit dans la section précédente, à partir des mots clés les plus fréquents et des réseaux de co-citations, que l'on a pu identifier sur chacun des corpus d'algorithme visualisés, les communautés thématiques les plus représentées. La Figure 23 montre chacun de ces graphes avec leurs communautés identifiées. La manière dont les réseaux peuvent être observés ici ne permet pas au lecteur de refaire lui-même tout ou partie de ce processus d'étiquetage des communautés. Pour cela, il trouvera en Annexe A l'ensemble de ces graphes reproduit à une taille plus adaptée pour ce type de travail.

Chaque réseau représenté dans la Figure 23 contient plusieurs communautés thématiques. Dans chaque corpus, une de ces communautés s'attache à des recherches théoriques ou fondamentales sur l'algorithme qui constitue le corpus lui-même. On les identifie dans chaque réseau par le nom du corpus indiqué en majuscules rouges.

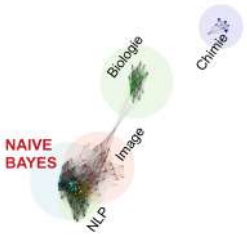
La distance entre les clusters dépend directement du nombre de liens entre deux communautés. Par exemple, le cluster de biologie dans le réseau des SVM (Figure 23f) est assez excentré dans la carte finale car il n'est connecté qu'à un autre cluster : la communauté des neurosciences, à travers deux liens seulement. Les neurosciences n'étant, par ailleurs, connectées à la communauté de médecine que par un seul lien. Cette dernière est par contre plus proche du cluster de recherche théorique (5 liens).

CHIMIE : Cette communauté traite de différents contextes d'utilisation d'outils informatiques et mathématiques en chimie. Les procédures d'apprentissage s'avèrent pertinentes notamment dans la chémométrie dont le but est d'obtenir le plus d'informations possible à partir de données chimiques, par exemple en construisant des *molecular descriptors* qui transforment l'information chimique en données numériques afin de permettre leurs traitements mathématique et informatique. Une des applications connue de ce processus se retrouve dans l'établissement des "relations quantitatives structure à activité" (en anglais, QSAR) qui visent à révéler les effets d'une structure chimique sur l'activité biologique ou la réactivité chimique. En ce sens, on retrouve au centre de ce groupe des journaux comme *Journal of chemical information and modeling*, ou *Journal of medicinal chemistry*. Les vertus analytiques des QSAR sont utilisées de plus en plus fréquem-

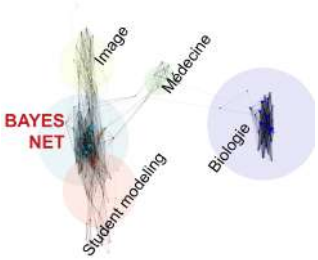
(a) cf. Figure 39



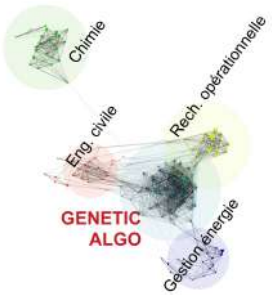
(b) cf. Figure 40



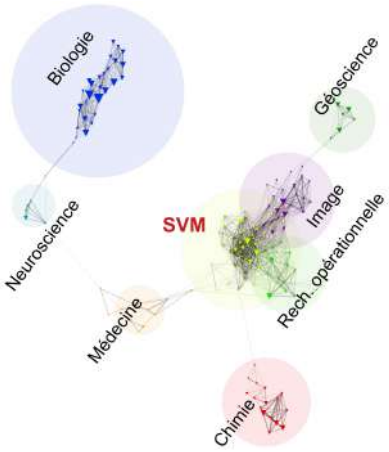
(c) cf. Figure 41



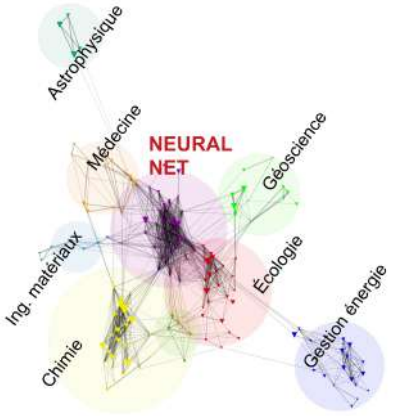
(d) cf. Figure 42



(e) cf. Figure 43



(f) cf. Figure 44



(g) cf. Figure 45

FIGURE 23 – Réseaux de co-citations des 150 journaux les plus cités pour chaque corpus

ment pour réglementer les substances chimiques. Par exemple, dans l'Union Européenne, depuis 2006, celles-ci font l'objet d'une réglementation commune (nommée REACH, pour *Registration, Evaluation, Authorization and restriction of CHemicals*) qui unifie, notamment grâce au QSAR, les procédures d'enregistrement, d'évaluation et d'autorisation des substances chimiques pour toute l'Europe.

BIOLOGIE : Cette thématique regroupe principalement des publications ayant trait à la biologie moléculaire, biologie des systèmes et de manière plus générale la bioinformatique, c'est à dire l'utilisation d'outils informatiques dans l'extraction et l'analyse d'information à partir de données issues de systèmes vivants. Cela inclut des revues spécialisées comme *Journal of Molecular Biology* et *Bioinformatics* et des journaux interdisciplinaires à fort impact comme *Science*, *Nature* ou PLOS. Un des usages importants de l'apprentissage dans ce contexte est de prédire certains attributs des protéines, leur localisation, leur repliement, leur connexion en réseau, etc. En génétique on retrouve l'usage de l'apprentissage statistique notamment pour la prédiction de l'expression des gènes, c'est à dire comment l'information héréditaire qu'ils contiennent peut produire telle ou telle molécule. La présence répétée de cette communauté dans presque tous les corpus montre bien comment la biologie est un champ d'application très friand des techniques d'apprentissage. Cependant sa place dans ces réseaux se situe souvent en périphérie de la recherche fondamentale des algorithmes mentionnés, c'est à dire, souvent, avec peu de références à la recherche qui traite des techniques considérées.

NEUROSCIENCE : Alors que la communauté des neurosciences est assez représentée sur le réseau du corpus *Machine Learning* (Figure 22), elle n'apparait dans les corpus d'algorithmes que pour les SVM. Les mots-clés font référence à l'imagerie par résonance magnétique (IRM) et le journal le plus fréquent, *Neuroimage*, laisse imaginer que les SVM sont souvent employés à des fins d'analyse d'image pour l'aide ou l'automatisation de diagnostic, notamment pour la maladie d'Alzheimer. Si cette communauté se distingue de celle du traitement de l'image qui est beaucoup plus importante et fréquente parmi les corpus, c'est qu'elle traite de sujets spécifiques. En observant les sujets centraux de ce groupe, on peut voir que les techniques d'apprentissage sont en partie utilisées pour leurs vertus d'intelligibilité. En effet, plusieurs études font état de l'usage des SVM pour extraire des variables pertinentes (*features extraction*) à partir de signaux extraits d'instruments de mesure comme les IRM, mais aussi d'électroencéphalographie (EEG). Par exemple, en cherchant à lier une certaine activité du cerveau avec une action demandée

au sujet de l'expérience, si on reconstruit les variables qui appuient cette prédiction, on peut en retour faire des inférences sur les zones d'activités neurologiques. Cet usage, qui s'appuie sur l'intelligibilité des procédures prédictives peut expliquer la spécificité des neurosciences et sa dissociation claire des communautés de traitement du signal.

ÉCOLOGIE : Cette communauté rassemble des études notamment sur les ressources hydrauliques (*Water ressource research*), la diversité et la préservation des espèces (*Diversity and Distributions*), les impacts environnementaux (*Global change biology*) autour de revues plus centrales comme *Ecology* et *Ecological Modelling* qui caractérisent ce groupe comme traitant d'écologie. Comme nous l'avons vu dans le chapitre précédent (§2.1), ce domaine d'application est un des contextes dans lequel les forêts aléatoires ont été formulées. Cela peut expliquer, au moins en partie, la forte présence de cette thématique dans ce corpus. Le seul corpus dans lequel cette communauté apparait aussi, dans une bien moindre mesure, est celui des réseaux de neurones. L'objet classique de ces procédures dans le contexte de l'écologie est double. D'abord la classification est utilisée afin de découvrir ou reconstruire les espèces lorsque l'on possède de nombreuses variables sur des observations d'un milieu naturel, comme nous l'avons montré pour la classification des fleurs Iris (§1.1.1.2). Ensuite, de nombreux travaux utilisent l'apprentissage comme technique de simulation, par exemple pour tester la survie de certaines espèces en fonction de différents modèles de l'avenir climatique.

IMAGE : La communauté du traitement d'image apparait dans les corpus bayésiens et celui des SVM. Dans chaque cas, elle se place très près de la recherche fondamentale concernant l'algorithme considéré. Cela s'explique en partie par le fait qu'à la différence des autres communautés, le traitement de l'image n'est pas une thématique applicative à proprement parler mais plutôt un ensemble de méthodes adaptée à l'analyse de données sous forme d'images, qui peuvent ensuite s'inscrire dans des applications tierces. Ainsi, cette communauté absorbe souvent des domaines d'applications qui ne développent pas assez de traits singuliers pour constituer une communauté propre. Dans ce sens, on retrouve des journaux représentant plusieurs domaines d'applications, par exemple *Lancet* pour la médecine ou IEEE *Transactions on Geoscience and Remote Sensing* pour les géosciences. Ce sont les mots-clés les plus fréquents (*computer vision*, *face/object recognition*, *image classification*) qui font ressortir la problématique dominante du traitement de l'image. On retrouve alors de nombreuses revues soit directement sur cette thématique (IEEE *Transactions on Image Processing*, *International Journal of Computer*

Vision) soit sur son implémentation dans un champ scientifique particulier (*Neuroimage*, IEEE *Transactions on medical imaging*).

GÉOSCIENCE : Cette communauté traite de l'usage des techniques statistiques nécessaires dans plusieurs domaines des géosciences, ou "sciences de la terre". Il s'agit en premier lieu de systèmes d'information géographique (en anglais *Geographical Information System*, GIS), c'est à dire de dispositifs informatiques qui intègrent, stockent, analysent et interagissent avec un grand nombre de capteurs. Extraire des informations pertinentes de ces mesures est sans doute l'application principale des procédures d'apprentissage. En ce sens, on retrouve de nombreux journaux qui traitent spécifiquement de télédétection (*remote sensing*), c'est à dire la captation de données à distance, comme par exemple *Remote Sensing of Environment* ou *International Journal of Remote Sensing*, en périphérie de journaux de géosciences plus généralistes comme *IEEE Transactions on Geoscience and Remote Sensing* ou *Engineering Geology*. Les mots-clés font ressortir quelques thématiques récurrentes au sein de cette communauté comme l'étude des glissements de terrain, ou des travaux autour du projet *Landsat*, premier programme spatial d'observation de la Terre à finalité civile.

MÉDECINE : La communauté médicale apparait dans plusieurs corpus mais de manière différente. En effet, dans le cas des réseaux neuronaux et bayésiens, et des SVM, il s'agit d'un petit nombre de journaux d'imagerie médicale (*IEEE Transactions on Medical Imaging*) notamment la radiologie (*Radiology*). Les mots-clés de ces groupes mettent en lumière des contextes d'aide au diagnostic notamment pour les cancers du sein et de la prostate, ou pour l'épilepsie. Dans ce cas, on peut imaginer des scénarios d'utilisations où on établit des modèles qui permettent de distinguer, par exemple, des images de tumeur maligne et bénigne et ainsi accompagner voire remplacer le diagnostic fait par le médecin. Dans le cas des arbres de décision, le cluster de recherche médical est beaucoup plus dense et touche un plus grand nombre de spécialités médicales comme la rhumatologie (*Arthritis & Rheumatology*), l'urologie (*Journal of urology*), la cardiologie (*Journal of the american college of cardiology*) ou la pédiatrie (*Pediatrics*), etc. Les mots-clés de ce groupe soulignent des usages orientés vers la prise de décision et l'analyse coût-bénéfice de celle-ci. Ainsi on retrouve ici le double rôle, décrit précédemment (§2.1), des arbres de décision comme algorithme d'apprentissage et comme formalisation et visualisation de systèmes experts.

NLP ET RECHERCHE D'INFORMATION : Cette communauté représente les utilisations des techniques d'apprentissage pour le traitement du langage naturel (en anglais *Natural Language Proces-*

sing, NLP). Cette communauté est spécifique au corpus des classifieurs naïfs bayésiens car ceux-ci s'avèrent particulièrement efficaces pour cette tâche (cf. §2.2). Comme nous l'avons vu précédemment, l'ambition de ce champ d'application est d'extraire l'information, la formater et l'agencer pour son utilisation à des tâches comme celles mises en lumière par les mots-clés : la classification de documents (spam, thématiques, etc), l'analyse de sentiment. Alors que bien des journaux présents dans ce groupe ont une appellation généraliste sur le traitement et la recherche d'information (*Information Processing & Management*, *Journal of the American Society for Information Science and Technology*), leur synopsis indique clairement leur intérêt fort pour l'analyse des contenus textuel ou web, et la structuration d'informations à partir de l'expression naturelle des auteurs et contributeurs de ces médias. Il s'agit donc d'une thématique un peu hétérogène qui combine le traitement spécifique du langage naturel et la recherche plus généraliste d'informations.

APPRENTISSAGE HUMAIN : Cette communauté apparaît uniquement dans le corpus des réseaux bayésiens et traite de la modélisation du comportement des apprenants et de leurs interactions avec des dispositifs informatiques, dans le but de susciter, accompagner et valider l'acquisition de connaissances. Ce domaine de recherche est généralement nommé *Environnements informatiques pour l'apprentissage humain* (EIAH) et fait écho à plusieurs initiatives apparues dans les années 70 pour utiliser les techniques naissantes d'IA pour la modélisation de l'intelligence humaine. Comme nous l'avons vu (§2.2), les réseaux bayésiens, du fait de leur usage explicite des relations de causalité, sont à la fois une procédure d'apprentissage statistique et une manière, comme pour les arbres de décisions, de représenter un comportement ou un système expert. Ce dernier usage prend d'autant plus de sens du fait de l'hypothèse des neurosciences que le cerveau humain effectue des inférences bayésiennes lors de l'apprentissage. Ainsi, cette communauté est un exemple emblématique de la façon dont différentes perspectives sur l'intelligence et l'apprentissage convergent au sein d'une même communauté qui s'intéresse autant à des questions d'ingénierie (*Knowledge engineering*), à la compréhension du processus interne d'apprentissage (*Psychological review*) et à leur simulation à des fins d'optimisation (*lecture notes in artificial intelligence*).

INGÉNIERIE CIVILE : Ce domaine d'application rassemble de nombreux champs de l'ingénierie civile autour de journaux centraux comme *Computers & Structures*, *Engineering optimization journal* et *Journal of mechanical design*. Il s'agit principalement d'applications dans le champ des infrastructures civiles (*Computer Aided Civil and Infrastructure Engineering*), de la gestion des ressources

hydrauliques (*Journal of Water Resources Planning and Management*), des transports (*Transportation Research Record*) et des environnements sonores (*Journal of Sound and Vibration*). Certains journaux cités ou mots-clés font ressortir des problématiques transversales, parmi lesquelles les tests de fiabilité (*Reliability Engineering & System Safety*), l'optimisation (*Engineering Optimization*), la topologie, etc. Cette communauté n'apparaît que dans le corpus des algorithmes génétiques et, comme nous l'avons vu (§2.3), profite probablement des capacités de cette famille de procédures d'apprentissage à optimiser des critères multi-objectifs et faire évoluer un modèle avec des données très imparfaites.

RECHERCHE OPÉRATIONNELLE : Cette communauté rassemble les usages des techniques d'apprentissage appliquées à différents scénarios de recherche opérationnelle, notamment autour de journaux centraux comme *European Journal of Operational Research* et *Computers & Operations Research*. De manière générale, ce domaine traite de l'utilisation d'outils statistiques et mathématiques pour appuyer des décisions dans des contextes industriels qu'illustrent des revues comme *The International Journal of Advanced Manufacturing Technology*, *Management Science*, *Journal of Intelligent Manufacturing*. Cette communauté apparaît dans les corpus des SVM et d'algorithmes génétiques, mettant en avant pour chacun des mots-clés qui illustre des domaines d'applications sensiblement différents. En effet, dans le cas des algorithmes génétiques, on retrouve des problématiques de gestion des stocks et du temps sur les chaines de production, alors que dans le cas des SVM émergent des applications comme l'évaluation des risques-clients (*credit scoring*), c'est à dire de la solvabilité d'un client ou d'un investisseur, et le diagnostic de leur défaillance potentielle.

GESTION DES ÉNERGIES : Cette communauté traite spécifiquement des usages de l'apprentissage artificiel dans la gestion des énergies et des systèmes énergétiques. Ainsi, autour du journal *IEEE Transactions on Power Systems* on trouve un certain nombre de publications sur la production énergétique, sa conversion (*IEEE Transactions on Energy Conversion*), sa livraison (*IEEE Transactions on Power Delivery*), son renouvellement (*Renewable Energy*). On retrouve donc ici un domaine spécifique d'application qui combine plusieurs problématiques traitées dans les autres réseaux comme la topologie, la gestion de ressources dans le temps, etc. Ce domaine particulier n'émerge cependant que dans les corpus des algorithmes génétiques et des réseaux de neurones.

INGÉNIERIE DES MATÉRIAUX : Cette communauté spécifique au corpus des réseaux de neurones rassemble quelques journaux autour des procédés mécaniques industriels de traitement des mé-

taux et autres matériaux. Ainsi on retrouve autour de *Journal of Materials Processing Technology* plusieurs journaux comme *Materials & Design*, *International Journal of Machine Tools and Manufacture* et *International Journal of Production Research*. Les mots-clés principaux qui émergent de ce groupe font ressortir quelques applications spécifiques des techniques d'apprentissage à ce domaine d'application comme la mesure de l'état de surface (*surface roughness*), ou l'extraction des propriétés mécaniques de certains matériaux (*mechanical properties*).

ASTROPHYSIQUE : Cette communauté rassemble des usages de techniques d'apprentissage dans le contexte de l'astrophysique, c'est à dire de l'étude de la physique et des propriétés des objets de l'univers (étoiles, planètes, etc). Ce groupe n'apparaît que dans les corpus des réseaux de neurones et des forêts aléatoires, à chaque fois comme une communauté très périphérique et rassemblant peu de journaux. Dans le cas des forêts aléatoires, il émerge des mots-clés relativement génériques à l'apprentissage (*data analysis*, *feature extraction*) qui laissent imaginer des usages assez variés dans le contexte général d'analyse de données. Dans le cas des réseaux de neurones, il s'agit principalement d'analyse d'image (*image processing*, *wavelet transform*).

À la différence des domaines d'intérêts référencés par WOS (§3.1), les thématiques qui émergent de l'analyse des réseaux de co-citations montrent la variété des domaines de recherche dans lesquels les techniques d'apprentissage sont mobilisées sans que cette observation soit biaisée par la catégorisation dominante en informatique ou ingénierie. Ainsi, on est plus à même d'observer la présence de chaque thématique et le fait qu'elle dépende, ou non, d'un ou plusieurs algorithmes.

La Tableau 10 résume ces éléments pour chaque thématique et permet de situer comment chacune est présente dans les différents corpus des algorithmes envisagés. On voit que 5 des thématiques sont présentes dans seulement un corpus. Il semble qu'il s'agisse de cas où les contraintes d'une thématique de recherche sont résolues par un unique algorithme. C'est le cas par exemple des neurosciences qui emploient les SVM à la fois pour leur performance en traitement d'images et en extraction de variables pertinentes. On trouve des cas similaires dans l'usage de la causalité dans les modèles bayésiens pour la modélisation des comportements d'apprenants, ou les vertus d'optimisation multi-objectifs des algorithmes génétiques et l'ingénierie civile. À l'inverse, on observe que 4 thématiques font appel à de nombreux algorithmes, plus précisément la biologie, la chimie, les géosciences et la médecine. Cela peut s'expliquer en partie par le fait que ces domaines rassemblent une variété d'usages qui s'expriment chacun avec des algorithmes différents. Une autre hypothèse possible est qu'au sein de ces thématiques se posent des problématiques d'op-

Domaine	Arbres de décision	Forêt aléatoire	Naïf bayes	Bayes net	Algo génétique	SVM	Neural Net
Biologie	✘	✘	✘	✘		✘	
Chimie		✘	✘		✘	✘	✘
Géoscience	✘	✘				✘	✘
Médecine	✘			✘		✘	✘
Image			✘	✘		✘	
Écologie		✘					✘
Astrophysique		✘					✘
Rech. op.					✘	✘	
Gest. énergie					✘		✘
NLP			✘				
Student model.				✘			
Eng. civil					✘		
Neuroscience						✘	
Ing. matériaux							✘

Tableau 10 – Présence de chaque algorithme par thématique de recherche

timisation plus génériques, par exemple de classification supervisée, qui peuvent ainsi être résolues par différents algorithmes.

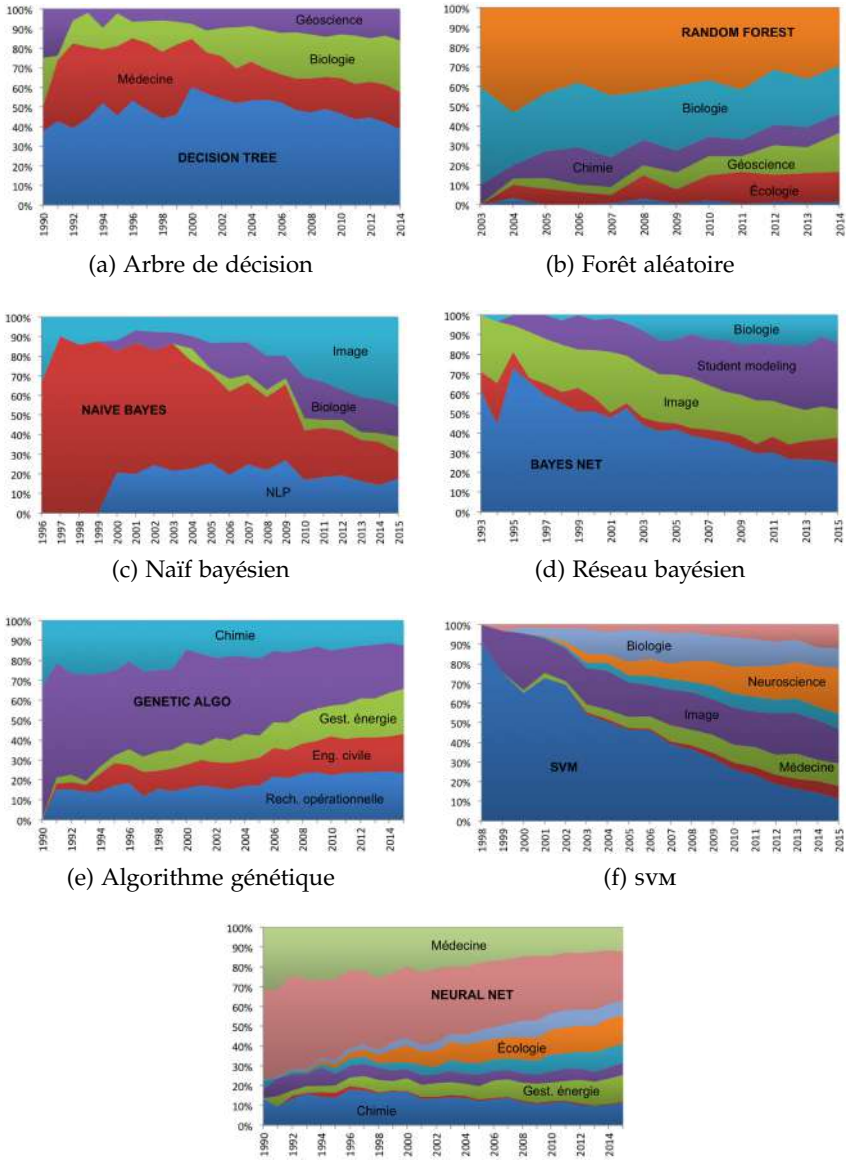
Le fait d'avoir identifié ces thématiques au sein des différents corpus d'algorithmes nous permet aussi de saisir leurs spécificités au sein du champ de l'apprentissage artificiel. Dans la section suivante on essaye de mieux caractériser la dynamique de ces différents champs applicatifs et de leur articulation à travers la trajectoire de leurs auteurs.

3.3.2 *Démographie des thématiques dans les communautés d'algorithmes*

L'analyse des réseaux de co-citations qui nous a permis de constituer le répertoire des thématiques de l'apprentissage artificiel dans le champ académique est le fruit d'une analyse statique, c'est à dire qui ne tient pas compte du temps. Cependant, une fois les clusters projetés sur une partie de la base de données, on peut réintroduire la variable temporelle *via* les dates de publications de chaque référence et ainsi observer l'évolution de chaque thématique dans le temps pour analyser le processus de structuration de chaque communauté d'algorithme.

Ainsi la Figure 24 nous permet d'observer comment, dans la plupart de ces communautés, la recherche théorique sur l'algorithme laisse progressivement une place de plus en plus grande aux applications thématiques. Si ce mouvement est très net pour les SVM dont les bornes temporelles de notre corpus permettent de capturer l'ensemble de la dynamique, il l'est moins pour les algorithmes plus anciens comme les approches bayésiennes, les réseaux de neurones et la programmation génétique. Cependant, la fenêtre de temps observable pour ceux-ci laisse observer une structuration similaire. Les arbres de décision et les forêts aléatoires font exception à cette observation, et la part de recherche théorique respective reste relativement stable et importante, se maintenant tout le long de la période considérée à plus de 30% de la production académique. Pour les forêts aléatoires, cela est d'autant plus étonnant que, comme pour les SVM, nous sommes en mesure de capturer de manière exhaustive l'existence académique de cette jeune famille de procédures d'apprentissage. Cela est peut-être dû à une activité de recherche plus intense sur les forêts aléatoires et de pistes très actives de recherche quant à l'amélioration de cette procédure. Dans le cas des arbres de décision, l'importance continue de la recherche théorique en la matière peut s'expliquer par la double identité de celle-ci comme technique d'apprentissage et comme modélisation des systèmes experts.

En ayant défini les thématiques de manière constante pour tous les corpus, on peut observer comment l'ensemble des thématiques se structurent dans le temps en prenant en compte la somme de tous les



(a) Arbre de décision (b) Forêt aléatoire

(c) Naïf bayésien (d) Réseau bayésien

(e) Algorithme génétique (f) SVM

(g) Réseau de neurones

FIGURE 24 – Démographie des thématiques dans chaque communauté d'algorithme

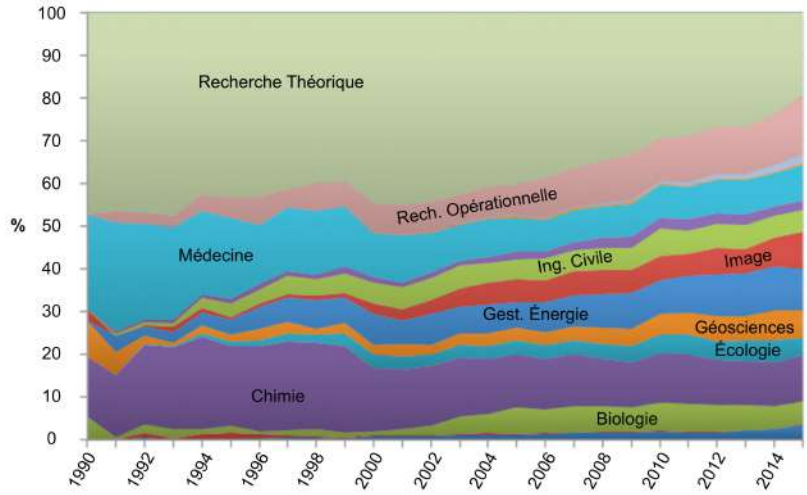


FIGURE 25 – Place de chaque thématique dans toutes les communautés d'algorithmes confondues

corpus d'algorithme. Ainsi, la Figure 25 permet de voir que la place de la recherche théorique est importante mais diminue sensiblement dans son ensemble, passant de moins de 50% de l'espace thématique en 1990 à un peu plus de 20% en 2015. Parmi les domaines d'application très représentés par ailleurs, on voit que la médecine et la chimie occupe une place importante mais sensiblement moindre que dans les années 90. Outre ces deux thématiques, la plupart des autres thématiques fortement représentées ne sont pas celles qui étaient présentes dans le plus grand nombre de corpus mais plutôt dans les corpus rassemblant le plus grand nombre de références. Ainsi, la recherche opérationnelle et la gestion des énergies occupent une place importante parce qu'elles sont présentes de manière significative dans les corpus les plus importants (SVM, algorithmes génétiques, réseaux de neurones).

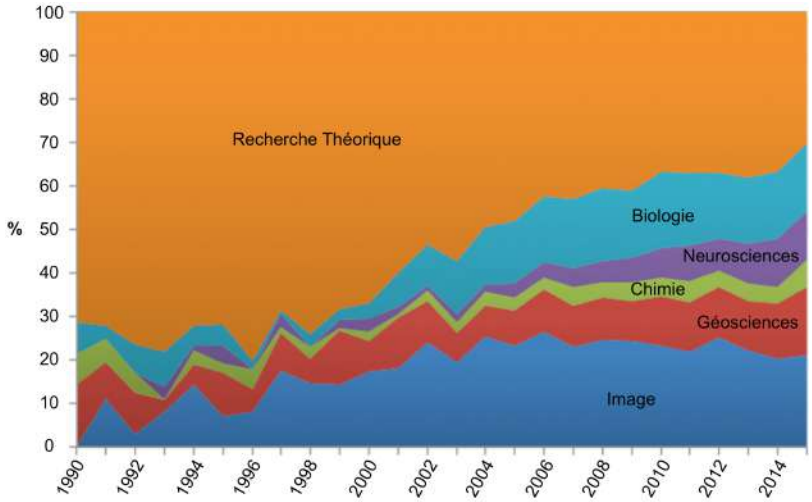


FIGURE 26 – Démographie des thématiques du corpus *Machine Learning*

Le tableau des thématiques de l'apprentissage que dresse la Figure 25 dépend de l'utilisation de certains algorithmes que nous avons définis en amont comme représentatif de la pratique de l'apprentissage artificiel. Si on revient sur le corpus des articles qui mentionnent le *Machine Learning* on peut probablement saisir des tendances différentes. Dans ce sens, la Figure 26 rend compte de la démographie des thématiques au sein du corpus *Machine Learning*. On peut observer que la part de la recherche théorique est beaucoup plus importante, passant de 70% d'occupation de l'espace thématique en 1990 à 30% en 2015, reproduit ainsi l'idée que si ce domaine de recherche se fonde sur une recherche fondamentale et plusieurs axiomes théoriques, il se structure autour d'applications qui légitiment son succès. Les domaines mis en valeur diffèrent aussi de la figure précédente en ce que sont mises en avant des applications qui font plus écho aux exemples historiques apparus dans le chapitre 2, comme le traitement de l'image et la biologie. Dans une certaine mesure, on peut considérer la différence entre les Figure 25 et Figure 26 comme la distance entre ce qu'incarne l'expression contemporaine du *machine learning* avec la pratique qui la précède et la définit historiquement. Dans ce sens, le renforcement des algorithmes plus récents permettent à des applications plus minoritaires ou inédites de prendre place dans ce champ de recherche.

3.3.3 *Distributions thématiques des auteurs*

Le fait de pouvoir observer les thématiques de manière transversale à tous les corpus des communautés d'algorithme nous permet d'étudier le comportement des auteurs à un degré plus fin d'analyse. L'ensemble de ces corpus rassemble 271,966 auteurs. De manière assez attendue, et comme le montre la Figure 27, la distribution du nombre de publications par auteur suit une loi de puissance, c'est à dire qu'un grand nombre d'auteurs publie peu et un faible nombre publie beaucoup, l'essentiel des auteurs ayant finalement publié un nombre de papiers assez distant de la moyenne générale. Notre intérêt se portant sur le potentiel déplacement des auteurs d'un algorithme à un autre ou d'une thématique à une autre, nous ignorons ici les auteurs n'ayant publié qu'une seule fois, portant ainsi le nombre d'auteurs observés dans cette section à 114,965 (42% du nombre total).

On considère qu'un auteur fait partie de plusieurs "communautés d'algorithme" s'il apparaît dans plusieurs des corpus décrits au début de ce chapitre (§3.1.1). La Figure 28 permet de visualiser comment, parmi les 45,956 auteurs ayant au moins publié dans deux de ces communautés (16%), quelles sont les communautés qui apparaissent le plus souvent ensemble. Plus formellement, on construit d'abord la matrice C de co-occurrences des algorithmes dans le par-

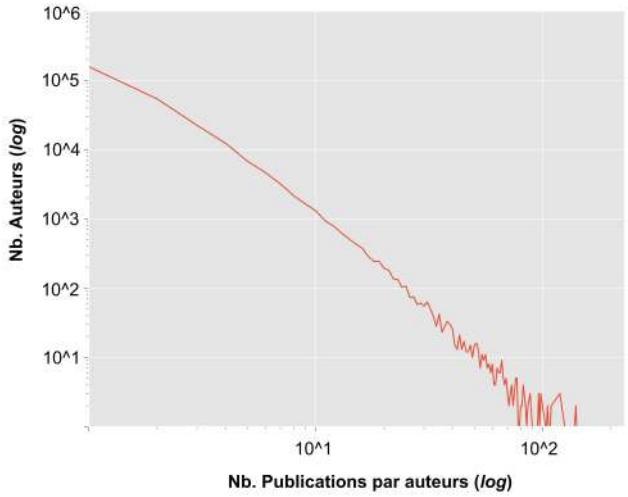


FIGURE 27 – Nombre d'auteurs par nombre de publications

cours des auteurs. Typiquement, si un auteur publie à la fois dans le corpus *Naive Bayes* et *Neural Net*, on rajoute 1 à l'entrée correspondante à ces deux algorithme dans la matrice C. On construit ensuite une matrice intermédiaire $\widehat{C}$ (cf. Équation 4) qui correspondrait à une répartition parfaitement aléatoire de l'ensemble des co-occurrences en ne contrôlant que la somme des lignes et des colonnes qui reste inchangée par rapport à la matrice originale. Enfin, pour chaque cellule de la matrice finale (cf. Équation 5) on calcule simplement le ratio entre la valeur observée des co-occurrences et celle qui serait produite par un comportement aléatoire des auteurs. On prend enfin le logarithme de cette valeur afin de la centrer sur 0. Ainsi, on obtient une *matrice de co-présence* des auteurs par communautés d'algorithmes qui montre combien deux algorithmes sont conjointement mobilisés par les mêmes auteurs : le score est positif, (cellule rouge) si les deux algorithmes sont - relativement aux autres couples d'algorithmes - souvent mobilisés par les même auteurs et négatif dans le cas contraire (cellule bleue).

$$\widehat{C}_{ij} = \frac{\sum_k C_{ik} \sum_l C_{lj}}{\sum_{kl} C_{lk}} \tag{4}$$

$$\text{Score} = \log\left(\frac{C_{ij}}{\widehat{C}_{ij}}\right) \tag{5}$$

Ainsi, si un auteur a publié dans le corpus *Naive Bayes*, il est fortement attendu qu'il ait publié aussi dans le corpus *Bayes Nets* par rapport à la probabilité qu'il publie dans un autre corpus au hasard. Dans ce cas, le lien est assez évident puisqu'il s'agit d'algorithmes

très proches, voir seulement de variantes d'une même tradition d'apprentissage statistique. On retrouve d'ailleurs une surreprésentation équivalente entre les arbres de décision et les forêts aléatoires qui partagent eux aussi un fort ancrage commun dans la tradition statistique. Mais on retrouve également des associations plus surprenantes : comme le lien très fort entre réseaux de neurones et algorithmes génétiques. On peut imaginer que le fait que ces deux algorithmes soient les plus anciens parmi ceux considérés ici et qu'ils soient tous les deux bio-inspirés augmente sensiblement la probabilité qu'un auteur qui utilise l'un puisse utiliser l'autre. Le lien fort entre les naïfs bayésiens et les arbres de décision peut, quant à lui, probablement s'expliquer à la fois par la simplicité et l'intelligibilité de la procédure et le succès de celles-ci. Enfin, on peut voir que les SVM co-occurrent de manière assez uniforme avec les autres algorithmes, et ce de façon quasiment toujours positive. Cela montre probablement que les SVM sont un choix qui dépend peu des autres traditions algorithmiques qu'on explore et qu'il est ainsi l'algorithme le plus compatible avec l'ensemble des familles du corpus.

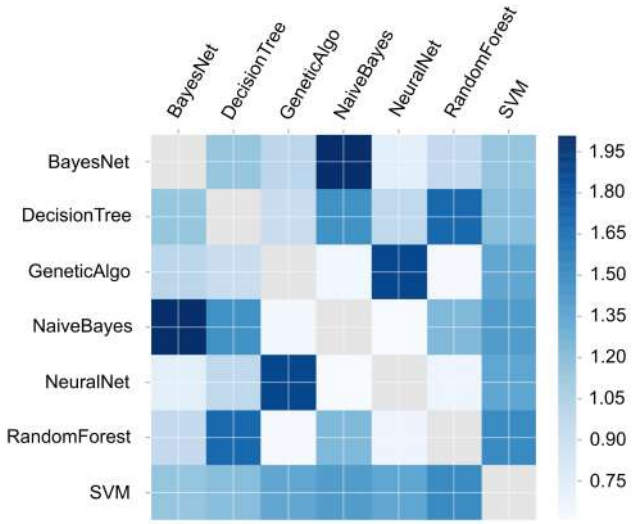


FIGURE 28 – Matrice de co-présence des auteurs par communauté d'algorithme

De la même manière que nous venons d'observer les sur/sous-représentations entre algorithmes d'apprentissage parmi les publications des auteurs, on peut observer celles des thématiques traitées. Par "thématiques" ou "domaines d'applications", on désigne les 14 groupes identifiés de manière transversale à tout les réseaux de co-citations et décrits dans la section §3.3.1, auquels on ajoute un groupe de "théorie" qui rassemble tous les clusters de recherche fondamentale sur les algorithmes eux-mêmes.

Ainsi, parmi les 84,903 auteurs ayant publié dans au moins deux domaines (31% du nombre total), si un auteur a publié dans la théma-

tique "apprentissage humain", il y a de fortes chances qu'il ait publié dans le groupe "image". Cette probabilité très forte s'explique probablement par le fait que le groupe "apprentissage humain" est une thématique propre à la pratique des réseaux bayésiens -corpus dans lequel était fortement représenté le groupe de traitement de l'image- . On retrouve une situation similaire pour expliquer la forte probabilité entre médecine et astrophysique, cette dernière thématique ne concernant que quelques publications, notamment dans le corpus des réseaux de neurones, où elle partage quelques liens seulement avec le groupe médecine. Il semble donc que les thématiques n'apparaissant que dans un faible nombre de corpus ont tendance a avoir un lien fort avec les thématiques apparaissant dans le même corpus. C'est le cas aussi par exemple du traitement du langage qui n'apparait que dans *Naive Bayes* et qui est fortement surreprésenté avec le traitement de l'image présent lui aussi dans ce corpus. De manière plus générale, cela montre qu'un algorithme commun pour traiter plusieurs thématiques semble pouvoir expliquer le déplacement d'un auteur de l'une à l'autre.

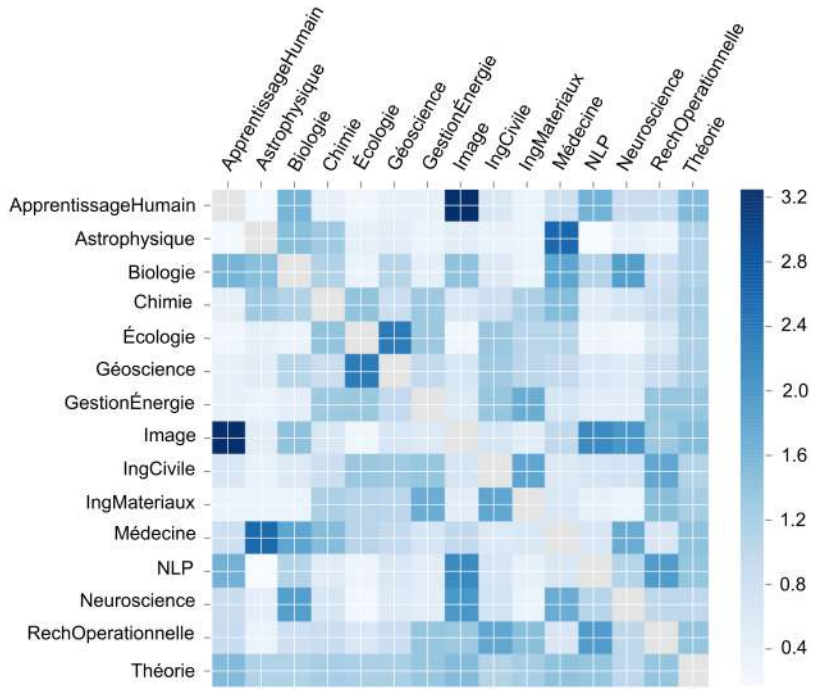


FIGURE 29 – Matrice de co-présence des auteurs par thématique

Surreprésentés dans une moindre mesure, on trouve des rapprochements thématiques plus naturels, c'est à dire plus proche de la manière dont s'identifient généralement les disciplines de recherches et désignent leurs disciplines connexes. Ainsi, les liens forts entre biologie et neurosciences, entre écologie et géosciences, entre médecine et biologie, semblent bien exprimer une logique de rapprochement thématique et s'écarter des opportunités de rapprochement algorith-

miques. On trouve aussi des liens forts entre des disciplines plus distantes thématiquement mais dont l'utilité réciproque dans le contexte de l'apprentissage fait sens. Ainsi, la forte probabilité de cooccurrence entre traitement de l'image et neurosciences s'explique, comme nous l'avons vu, par l'usage de l'apprentissage dans les neurosciences principalement pour l'analyse d'IRM, d'EEG, etc. Enfin, on voit que la pratique théorique de l'apprentissage est représentée de manière homogène avec toutes les autres thématiques, avec un rapprochement toujours légèrement positif.

De manière plus générale, les Figure 28 et Figure 29 nous ont permis d'observer le déplacement des auteurs présent dans nos corpus selon deux coupes transversales, respectivement celle des domaines thématiques, et celle des algorithmes. Dans chaque cas, nous avons vu que des logiques à la fois endogènes et exogènes pouvaient expliquer la constitution de forts liens entre plusieurs thématiques ou algorithmes. Ainsi, il semble que dans la carrière d'un chercheur présent dans nos corpus, son déplacement, puisse, entre autre, être le fruit de logiques thématiques et algorithmiques.

Si ces deux figures nous ont permis de décrire quelques dynamiques de ces logiques, il est difficile de juger de la prédominance de l'une ou de l'autre comme déterminant du déplacement des auteurs. La Figure 30 permet d'observer la distribution du nombre de thématiques (ligne rouge) et du nombre d'algorithmes (ligne bleue) explorée par les auteurs. Il apparait nettement qu'il est plus probable qu'un auteur explore plusieurs thématiques plutôt que différents algorithmes. En effet, alors que la courbe des algorithmes se présente en une loi de puissance classique, la courbe des thématiques présente le caractère singulier qu'il est bien plus probable, parmi les auteurs ayant au moins publié deux fois, que ce soit dans deux thématiques différentes plutôt qu'une. Il semble donc que parmi les auteurs que nos corpus permettent d'observer, la pratique d'un ou plusieurs algorithmes soit moins changeante et donc plus caractéristique des comportements que la pratique des thématiques. Cette observation va dans le sens d'une cohérence d'un domaine de recherche du *Machine Learning* en tant que tel, où la pratique et l'exploration de procédures d'apprentissage transcende celle des thématiques abordées avec ces techniques.

Nous tirons cette conclusion de l'étude faite dans ce chapitre de traces des chercheurs faisant références à plusieurs techniques d'apprentissage. Ces données ont l'avantage d'être représentatives car résultant de plusieurs décennies d'activité de publication. Cependant ces données ne sont pas forcément représentatives d'efforts plus récents et de pratiques moins académiques qui sont apparues en apprentissage artificiel ces dernières années. C'est ce domaine d'activité plus caractéristique des dynamiques contemporaines que le chapitre 4 tente de cerner.

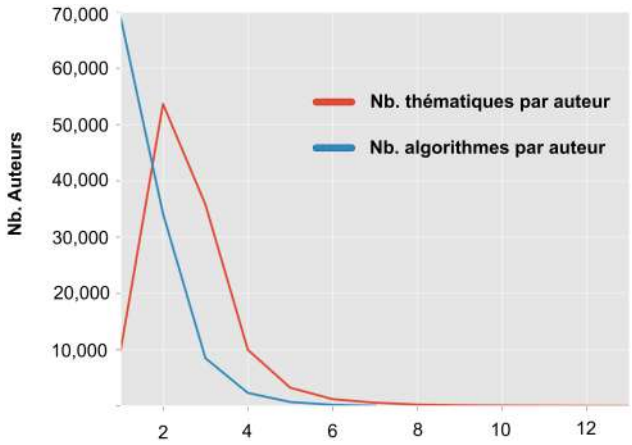


FIGURE 30 – Nombre d'auteurs par nombre de domaines et par nombre d'algorithmes

RÉSUMÉ DU CHAPITRE 3

Ce chapitre vise à rendre compte des communautés qui se construisent autour des principaux algorithmes d'apprentissage, à partir de données représentant l'activité de la recherche académique. Ainsi, en extrayant plusieurs corpus de *Web of Science*, on parvient à décrire plusieurs caractéristiques de ces communautés en termes de nombre d'auteurs, de leur renouvellement, pays d'origine, domaines d'intérêts, etc.

Dans un deuxième temps, on a décrit des méthodes d'analyse de réseau qui nous permettent d'observer la structure des co-citations des journaux académiques les plus présents dans chaque corpus. Leurs structures hauts-niveaux, révélées par un algorithme de regroupement, nous permet d'observer les communautés thématiques explorées par chaque algorithme d'apprentissage envisagé.

Ces communautés thématiques sont transversales à de nombreux corpus. On peut alors en faire la démographie et montrer comment les auteurs ont tendance à explorer davantage de thématiques que d'algorithmes et ainsi caractériser l'apprentissage artificiel comme un opportunisme méthodologique.

4

UN APERÇU DE QUELQUES USAGES CONTEMPORAINS

SOMMAIRE

Le chapitre précédent nous a permis de rendre compte de certains usages de l'apprentissage artificiel dans le champ académique, dont on a observé la séparation entre les démarches théoriques et applicatives. Dans le présent chapitre, nous déplaçons notre regard hors du champ académique pour concentrer notre attention sur la dimension applicative de l'apprentissage et sur les efforts d'ingénierie qui y sont déployées. En complétant notre étude de la sorte, on confirme bien notre intérêt pour les conditions de production des dispositifs d'apprentissage, délaissant ainsi de nombreux autres lieux pertinents à la caractéristation de ce domaine d'activité technologique. Comme autre point d'intérêts auquel on aurait pu penser pour analyser la dimension applicative, il y a bien évidemment les interactions entre les utilisateurs qui bénéficient de services quotidiens de plus en plus nombreux intègrant ces techniques d'apprentissage. Les entreprises qui cherchent à combiner techniques, bases de données et communautés d'utilisateurs et ont recours aux techniques d'apprentissage jettent probablement les bases de nombreux rapports de forces à venir. Pour l'heure donc, nous nous intéressons aux communautés d'ingénieurs qui font un usage plus distant des algorithmes d'apprentissage, en ce sens qu'elles se libèrent souvent de l'obligation d'avoir une compréhension fine de leurs fonctionnements, et s'intéressent davantage aux contraintes et besoins du développement de leur projets.

Afin de pouvoir saisir des traces de ces comportements, nous nous sommes focalisé sur plusieurs services en ligne de collaboration et d'interactions entre amateurs et professionnels de l'apprentissage et de ses disciplines connexes. La principale raison de ce choix est que ces plate-formes comportent suffisamment de données sur leurs utilisateurs pour que nos observations puissent prétendre à une certaine forme de pertinence statistique et de représentativité. En ce sens, on exploite donc les traces d'utilisateurs des sites de questions-réponses afin de voir dans quels contextes la notion de l'apprentissage et de ses algorithmes interviennent (§4.1). Dans un deuxième temps, nous nous intéressons à une plate-forme de compétitions de *machine learning* afin d'analyser l'impact de la nature des compétitions sur les choix méthodologiques des participants (§4.2). Plusieurs autres terrains auraient pu faire l'objet d'observations tout aussi légitimes, comme par exemple des lieux de discussion centraux pour ces communautés comme *Reddit*, certaines *mailing-list*, les forges de code comme *Github*, etc.

4.1 STACKEXCHANGE

Après avoir rapidement présenté le réseau des sites *Stackexchange* et les données qui peuvent en être extraites (§4.1.1), nous analyserons la présence de l'apprentissage dans l'ensemble de ces sites et en dégagerons ainsi quelques observations préliminaires sur la nature des ces techniques dans le contexte d'une plate-forme de questions-réponses (§4.1.2). En se concentrant sur les sites les plus pertinents, nous étudierons ensuite comment chacun représente un contexte particulier de l'apprentissage à partir des réseaux de cooccurrences des mots-clés les plus fréquemment associés à la mention du *machine learning* (§4.1.3). De cet ensemble de mots-clés, on extrait ceux qui représentent l'usage d'algorithmes spécifiques afin de pouvoir analyser comment ils apparaissent conjointement dans le parcours des utilisateurs. On pourra ainsi émettre plusieurs hypothèses sur les logiques qui dominent leurs choix et les comparer avec celles observées dans le champ académique (§4.2.3).

4.1.1 *Présentation du réseau Stackexchange*

Stackexchange [1] est un ensemble de forums permettant à leurs utilisateurs de poser des questions et d'y répondre. Chaque site est une déclinaison thématique du site pionnier *Stackoverflow*, créé en 2008, qui rassemble en 2016 environ 12 millions de questions et 6 millions d'utilisateurs autour de la pratique de la programmation et de l'in-

1. http://stackexchange.com/sites

formatique. Aujourd'hui, on trouve plus de 150 sites traitant de sujets aussi variés que le jardinage, l'islam, le latin, le travail du bois, etc. Cependant, la plupart des sites ayant rencontré le succès le plus notable sont ceux qui traitent de sujets autour de l'informatique, la science, les jeux vidéos et le web.

Les sites ont un système de modération élaboré, avec divers systèmes de badges, de médailles, de points, permettant d'accéder à certains privilèges de modération pour pouvoir valoriser ou pénaliser la visibilité d'une question ou d'une réponse, fermer une discussion, la mettre en exergue, etc. La modération sur ces sites est généralement considérée comme conservatrice ou sévère car elle favorise la suppression de toutes les questions imprécises, vagues, d'opinion. La volonté affichée des fondateurs [125] de ce réseau est d'exclure toutes les questions inutiles ou mal formulées afin que chacune pose un problème susceptible d'être résolu. Le corollaire d'un tel système de modération est que, de manière générale, ces sites sont très bien structurés, autour de mots-clés aussi variés que précis.

Depuis sa création le réseau stackexchange a une politique d'ouverture de ses données très affirmée et permet à tout un chacun de disposer de l'ensemble des données anonymisées de tous les sites depuis leur création [2]. Ces sites représentent donc une opportunité de recherche pour observer et analyser les interactions des utilisateurs dans un contexte thématique particulier. L'ensemble des observations et analyses ont été réalisées à partir des données mises à disposition par Stackexchange le 1er mars 2016.

Bien que relativement récentes, les données du réseau de sites Stackechange ont déja fait l'objet de nombreuses études. Certaines s'intéressent aux utilisateurs, comme Wang, Lo et Jiang [144] qui étudient les interactions entre les développeurs sur Stackoverflow, Vasilescu, Capiluppi et Serebrenik [142] qui analysent l'impact du genre dans celles-ci. Bosu et al. [16] se concentre sur les processus de réputation mis en perspective par les nombreux systèmes de récompense que proposent les sites. D'autres études se concentrent davantage sur le contenu, comme par exemple l'évaluation de la qualité des questions [83], la difficulté du problème posé [43] ou les mots-clés utilisés pour les étiqueter [126].

4.1.2 *Identifier les sites pertinents*

Afin d'identifier les sites pertinents à l'étude de l'usage de l'apprentissage artificiel, nous avons sélectionner tous les sites qui possèdent au moins une question référencée avec le mot-clé "machine-learning".

2. https://archive.org/details/stackexchange

Le Tableau 11 montre la liste de ces sites en indiquant pour chacun le nombre de questions concernées, ainsi que le pourcentage de toutes les questions du site que celui-ci représente.

Nom du site	**Accroche du site**	**Nb. tag**	**% site**
Stack Overflow	Programmeurs enthousiastes et professionnels	9095	0.03%
Cross Validated	Personnes intéressées par les statistiques, l'apprentissage artificiel, l'analyse, l'exploration et la visualisation des données	4937	3.27%
Data Science	Professionnels de la science des données, spécialistes en apprentissage automatique et ceux intéressés par le domaine	742	13.16%
Mathematics	Personnes intéressées par l'apprentissage des mathématiques à tout les niveaux, et les professionels dans les domaines concernés	652	0.04%
Computer Science	Étudiants, chercheurs et praticiens des sciences informatiques	383	1.21%
Theoretical Computer Science	Théoriciens informatique et chercheurs de domaines concernés	189	0.94%
Programmers	Programmeurs professionnels intéressés par des questions conceptuelles sur le développement logiciel	73	0.04%
Computational Science	Scientifiques qui utilisent des ordinateurs pour résoudre des problèmes scientifiques	68	0.54%
Signal processing	Praticiens des l'art et la science du traitement du signal, de l'image et de la vidéo	71	0.35%
Quantitative Finance	Professionnels et académiques de la finance	41	0.26%
Mathematica	Utilisateurs de Mathematica	67	0.08%
Code Review	Revue de code par les pairs programmeurs	46	0.05%
Robotics	Ingénieur robotique professionnel, amateur, chercheur et étudiants	25	0.39%
Open Data	Développeurs et chercheurs intéressés par les données ouvertes	64	1.26%
Software recommendations	Personnes cherchant des recommandations de logiciel spécifiques	18	0.10%
Geographical Information Systems	Cartographes, géographes et professionnel des GIS	8	0.01%
Meta Cross Validated	Discussion autour du site *Cross Validated*	5	0.17%
Ask Patents	Personnes intéressées par l'amélioration et la participation au système des brevets	4	0.06%
History of Science and Mathematics	Personnes intéressées par l'histoire et les origines de la science et des mathématiques	2	0.09%

Tableau 11 – Presence de l'apprentissage artificiel sur les site du réseau *Stackexchange*.

Lorsque une personne se rend sur un des nombreux site de *stackexchange*, on peut postuler qu'elle sélectionne ce site en fonction du

domaine d'expertise le plus adapté à sa question. Lorsque un utilisateur utilise le mot-clé "machine-learning" pour étiqueter sa question, il discrimine son problème dans ce domaine d'expertise. Ainsi, le nombre d'occurrences du mot-clé qui représente "apprentissage artificiel" dans un site donné, comparé aux autres sites, peut être interprété comme l'importance relative du domaine d'expertise du site en question vis-à-vis des problèmes posés par les méthodes d'apprentissage. On retrouve à cette place les domaines de connaissance déjà amplement présentés dans les chapitres précédents, à savoir l'informatique, principalement pour des questions d'implémentation (*Stackoverflow*), et les statistiques (*Cross-validated*). Dans une moindre mesure, on trouve le site *Data Science* ouvert plus récemment (2014) et qui rassemble des questions plus généralistes abordant le traitement des données de manière générique. Le site *Mathematics* aborde les différents champs des mathématiques sollicités par l'apprentissage (algèbre linéaire, optimisation). On voit ensuite plusieurs déclinaisons de l'usage de l'informatique, notamment avec un aspect plus scientifique et donc probablement moins tourné vers l'implémentation et davantage vers des questions plus abstraites ou fondamentales liées à l'apprentissage (*Computer science*, *Theoretical computer science*, etc). Enfin, à bien moindre échelle, on trouve divers domaines d'application de l'apprentissage déjà aperçus lors de l'étude des données académiques de WOS, comme le traitement du signal et les GIS. À ce même titre, on voit apparaître de nouveaux usages comme l'analyse quantitative en finance et la robotique.

À l'inverse du nombre brut d'occurrence du mot-clé "machine-learning", le ratio de ce nombre avec l'ensemble des questions du site représente la part de l'apprentissage dans la thématique du site, c'est à dire à quel point cette discipline est caractéristique du domaine d'expertise que représente la thématique du site. Dans ce sens, en comparant les deux sites qui attirent le plus de questions sur l'apprentissage, il est clair que la part de l'apprentissage dans l'informatique est insignifiante par rapport à celle qu'elle occupe en statistiques. De plus, si l'on considère que sur un forum il n'est pas nécessaire d'indiquer ce mot-clé pour discriminer des questions précises sur cette technique, on peut considérer que la part du site *cross-validated* consacré à l'apprentissage est bien supérieure aux 3.27% que ce ratio laisse entrevoir. Néanmoins, ce ratio est le plus important pour le site *data science* où le mot-clé référence presque 15% des questions, montrant, comme nous l'avons vu dans le chapitre 1, à quel point l'apprentissage artificiel est constitutif de la récente appellation "science des données".

Au vu des ces éléments descriptifs des différents sites de *Stackexchange* traitant d'apprentissage automatique, il semble que les quatre premiers éléments de la liste du Tableau 11 permettent à la fois de constituer un corpus de taille raisonnable (notamment en vue de

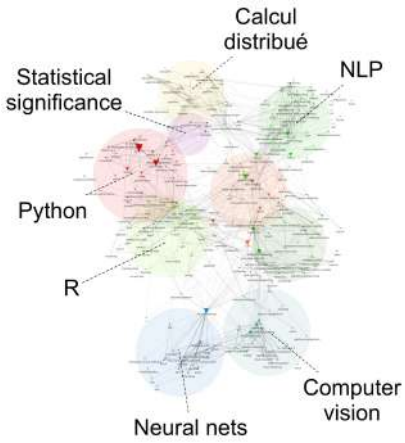
l'analyse des profils de occurrences de leurs mots-clés) et de permettre d'analyser des domaines : sensiblement définis par l'apprentissage (*Cross-validated*, *Data Science*), ou concernés de façon plus incidente par rapport à l'ensemble des questions échangées (*Stackoverflow*, *Mathematics*).

4.1.3 *Réseaux de cooccurrence de mots-clés*

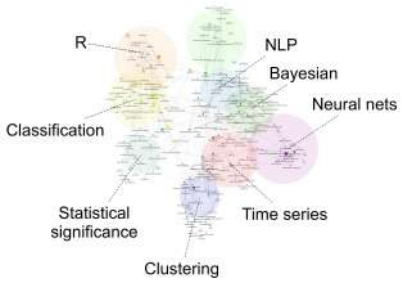
Afin de pouvoir analyser les sites sélectionnés, on a fait appel aux mêmes méthodes que celles sollicitées pour l'analyse des réseaux de co-citations à partir des données de WOS (cf. §3.2.2). Ici, pour un site donné, nous considérons chaque question référencée par le mot-clé "machine-learning". On retire ce mot-clé commun à toutes les questions sélectionnées, et, pour chacune, on créé un lien entre tous les mots restant. Enfin, on construit le réseau en retenant les liens qu'on observe avec une certaine fréquence dans l'ensemble des questions. La Figure 31 permet d'observer la structure haut-niveau de chaque réseau, révélée par l'algorithme de regroupement. Les groupes qui en émergent sont labellisés par une étiquette générale des mots-clés présent dans le groupe en question, appréciation dont le lecteur peut juger de la pertinence en observant les figures reproduites à plus grande échelle dans l'Annexe C.

Une première observation générale est que la structure de ces réseaux est beaucoup moins claire que celles des réseaux de co-citations académiques, où l'on pouvait observer davantage de niches clairement distinctes les unes des autres. Certes, dans le cas présent on observe des réseaux sémantique alors que l'analyse du chapitre précédent reposait sur des réseaux de co-citations. Il ne s'agit donc pas entièrement des mêmes objets et ceux-ci peuvent varier dans leur structure par ce simple fait. Néanmoins, on voit que chaque groupe entretient de nombreux liens avec la plupart des autres. La densité des liens entre les nœuds d'un même groupe justifie fragilement son existence, à tel point que les mots-clés de certains groupes ne permettent pas de leur attribuer un label. Cette absence de structure claire dans les réseaux est, en elle-même, un résultat qui désigne une communauté de savoir qui, dans les contextes des sites de questions-réponses, est assez homogène et peu divisée par les thématiques qui la composent. En les comparant avec les réseaux académiques, on peut envisager que les pratiques applicatives, d'implémentations, d'ingénierie de l'apprentissage artificiel sont probablement moins segmentées que les pratiques de recherche.

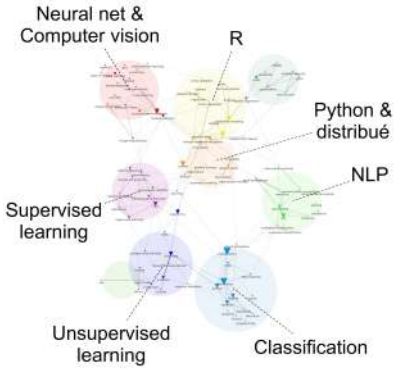
Les communautés qui émergent de ces réseaux sont de natures diverses. On trouve des langages de programmation dont la présence caractérise bien le site envisagé. En effet, un groupe spécifique au-



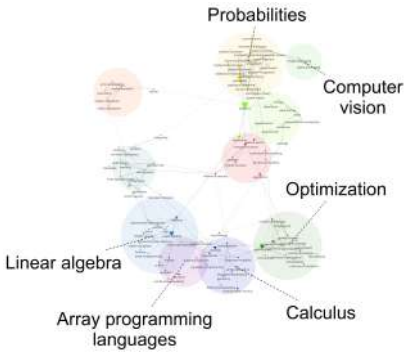
(a) **Stackoverflow** - Figure 46



(b) **Cross-validated** - Figure 47



(c) **Data Science** - Figure 48



(d) **Mathematics** - Figure 49

FIGURE 31 – Réseaux de cooccurrences des mots-clés de plusieurs sites Stackexchange.

tour de R, qui se présente comme un langage dédié à l'informatique statistique [3], est présent dans tous les corpus sauf celui du site *Mathematics*, allant ainsi dans le sens de l'ancrage de la communauté de l'apprentissage dans celle des statistiques, aussi bien dans son ingénierie que dans son implémentation. Le langage Python, qui a une visée bien plus généraliste, ne constitue une communauté à part que dans le site *Stackoverflow*. Le site *Mathematics* fait ressortir des langages d'un type particulier, qui mettent l'accent sur le traitement des matrices, comme Matlab, Octave ou Julia. La présence de cette communauté fait particulièrement sens aux abords de celles dont elle est très proche, l'algèbre linéaire et le calcul.

Les deux domaines d'applications qui émergent en tant que communautés sont le NLP (*text classification, analysis, processing, mining*) et le traitement de l'image (*object, image, gesture recognition, computer vision*). Comme nous l'avons vu dans le chapitre précédent, il s'agit des deux domaines qui caractérisent les plus récents développements de l'apprentissage. Les mots-clés ne font aucune mention des autres domaines importants rencontrés dans la sphère académique comme la médecine, la chimie, la recherche opérationnelle. Cela ne signifie pas forcément que ces domaines ne sont pas ceux que les questions essayent de traiter, mais que les données que ces domaines proposent n'impliquent pas de défis particuliers au point de constituer une communauté distincte. Ainsi, là où par exemple, dans le contexte académique, on pouvait voir les neurosciences tantôt être annexées par la communauté de traitement de l'image, tantôt en être indépendantes, ici on observe que l'usage des questions-réponses ne font quasiment pas mention de leurs contextes d'applications.

Parmi les communautés représentant une famille d'algorithmes en particulier, on ne trouve presque que les réseaux de neurones (*deep learning, convolutionnal nets, autoencoder*). On peut expliquer cette présence dans tous les corpus excepté celui de *Mathematics* par le fait de deux facteurs conjoints. Tout d'abord, comme nous l'avons déjà vu à plusieurs reprises, il y a un intérêt fort et croissant depuis 2012 pour ces techniques, que ce soit sous l'empire de l'appellation "deep learning" ou "neural networks". D'autre part, cette famille d'algorithmes nécessite beaucoup d'éléments de configuration spécifiques aux problèmes que l'on cherche à résoudre et en ce sens il est normal de trouver, outre sa grande popularité, beaucoup de questions qui l'identifie de manière précise au point de constituer une communauté.

À défaut de trouver beaucoup d'algorithmes directement représentés par les communautés mises en lumière par ces réseaux, on retrouve beaucoup plus fréquemment des communautés rassemblant des types d'objectifs et de contraintes que certaines familles d'algorithmes traitent efficacement. Dans ce sens on trouve, la typologie

3. "The R Project for Statistical Computing", https://www.r-project.org/

des types d'apprentissage décrite dans le premier chapitre (§1.1.1.3), qui distingue l'apprentissage supervisé (*classification*, *predictive-models*, *sampling*) de l'apprentissage non-supervisé (*clusering*, *distance-functions*, *outliers*). On ne retrouve néanmoins leur mention conjointe que dans le corpus *Data Science*. Une distinction proche que l'on retrouve davantage est celle entre *clustering*- l'application la plus courante de l'approche non-supervisée - et la *classification* - l'application la plus courante de l'approche supervisée -. Enfin un dernière communauté qui aborde les contraintes de certains algorithmes est celle des traitements de séquences temporelles (*time series*) qu'on retrouve uniquement dans le corpus *cross-validated*.

Si ces réseaux nous ont permis de diversifier l'observation des contextes sémantiques dans lesquels apparait le terme "machine-learning", on ne voit que très peu apparaître les algorithmes mis en lumière dans l'étude précédente issue du contexte académique. Pour cela, la section suivante vise à observer la place des mots-clés désignant les algorithmes observés jusqu'à maintenant dans cette thèse.

4.1.4 *Coprésences des algorithmes*

Si le mot-clé "machine-learning" permet de saisir un certain nombre de questions ayant trait à ce sujet, la mention d'algorithme précis nous permet d'explorer des questions qui dépassent ce premier domaine et de recouper des observations avec celles faites dans le chapitre précédent sur les communautés d'algorithme dans la sphère académique. Pour chaque site *stackexchange* considéré, on trouve plusieurs mots-clés qui peuvent faire référence à un même algorithme. Pour cela, on décrit en annexe la liste des mots-clés associés à chaque algorithmes considéré (§B.1). Tout simplement, il s'agit des mots-clés qui peuvent être associés exclusivement à un algorithme. Il s'agit le plus souvent de singuliers et pluriels, de variantes de nomination (par exemple *genetic-algorithm* et *evolutionnary-algorithm*), ou de la mention d'éléments spécifiques à un algorithme (par exemple CART pour les arbres de décisions). Le faible nombre de ces mots-clés sur les sites *Data Science* et *Mathematics* écarte ces sites de cette analyse et nous invite à se concentrer uniquement sur les sites *Stackoverflow* et *Cross-validated*, qui au vu du Tableau 11 semblent rassembler le cœur de l'activité autour de l'apprentissage artificiel.

La Figure 32 permet d'apprécier la présence de chaque algorithme dans les deux sites sélectionnés, exprimée en nombre d'occurrences. Tout d'abord, on voit que quatre de ces algorithmes (arbres, forêts et approches bayésiennes) ont sensiblement le même nombre d'occurrences dans les deux sites. Cette présence comparable peut exprimer le fait que ces algorithmes ne sont pas spécifiquement repré-

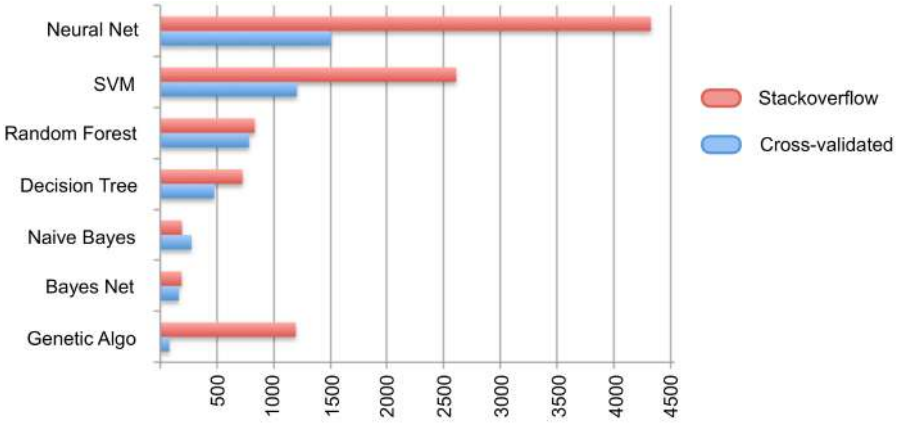
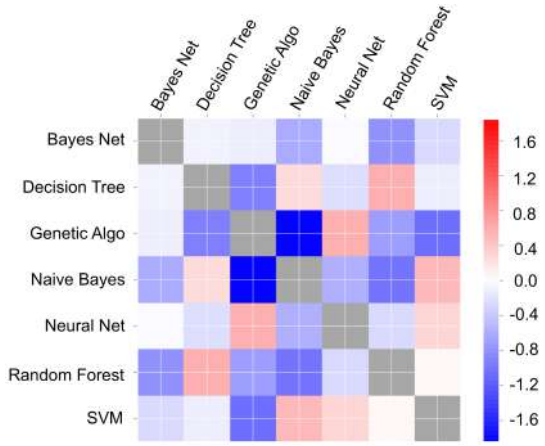


FIGURE 32 – Nombre de mots-clés par algorithme sur *Cross-validated* et *Stackoverflow*

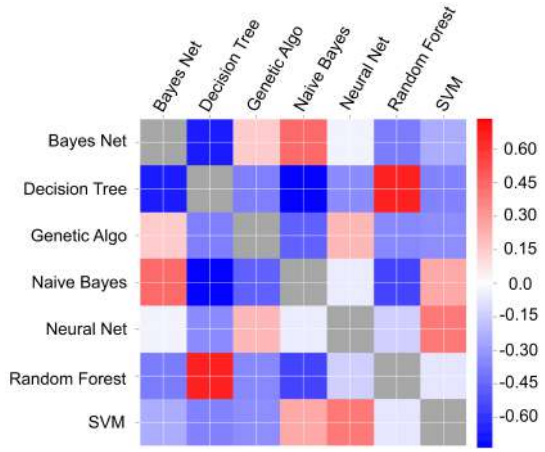
sentatifs de questions d'ordre statistique ou d'ingénierie, et se répartissent de manière assez équivalente entre les deux domaines représentés par les sites. *A contrario*, Les réseaux de neurones et les SVM se répartissent de manière nettement plus déséquilibrée d'un site à l'autre et on retrouve deux fois plus d'occurrences dans stackoverflow que dans cross-validated. On retrouve donc pour ces algorithmes la même disparité que pour l'apprentissage en général. Enfin les algorithmes génétiques sont assez fréquents sur le site *stackoverflow* avec un nombre d'occurrences comparable aux forêts aléatoires, mais cependant presque inexistants sur le site *cross-validated*. Cela valide l'hypothèse faite précédemment que si les algorithmes génétiques se sont véritablement identifiés à l'apprentissage artificiel, ils ne posent pas ou plus de questions particulières au champ des statistiques.

Cette observation comparée du nombre brut d'occurrences de chaque algorithme nous permet de capturer l'usage qui en est fait. Nous nous proposons maintenant de changer de perspective pour nous intéresser aux pratiques individuelles : est-ce que les utilisateurs de ces forums font mention de plusieurs algorithmes ? Lesquels sont « compatibles » ou « incompatibles » les uns avec les autres du point de vue des usages des internautes sur ces sites ? Afin de révéler la structure de ces relations entre algorithmes on utilise la même méthode que pour la distribution thématique des auteurs au sein des clusters de citation du chapitre précédent (§3.3.3), afin d'obtenir la matrice de coprésence des auteurs par algorithme, illustrée par la Figure 33. Le lien entre deux algorithmes est donc tissé par le fait qu'il co-occure dans l'historique des questions posées par un même utilisateur. Dans les rares cas où deux algorithmes sont mentionnés dans la même question, celle-ci est considérée deux fois.

Pour le site *stackoverlfow*, la Figure 33a permet donc de visualiser comment les auteurs ont tendance à explorer certaines combinaisons d'al-



(a) Stackoverflow



(b) Cross-validated

FIGURE 33 – Matrices de coprésence des utilisateurs par algorithme sur *Stackoverflow* et *Cross-validated*.

gorithmes plutôt que d'autres. Parmi les 6,967 utilisateurs considérés sur *stackoverflow*, seulement 511 (7%) ont posé des questions sur au moins deux algorithmes, et sont donc représentés dans la matrice de coprésence. De la même manière, sur le site *cross-validated*, 384 (13%) parmi les 2,758 utilisateurs considérés sont représentés par la Figure 33b. Le ratio beaucoup plus élevé pour le site spécialisé en statistique semble indiquer que cet ancrage disciplinaire permet de naviguer de façon plus fluide entre les différents algorithmes.

Dans le cas de stackoverflow, on voit que la cellule la plus intense de la matrice représente une répulsion forte entre l'usage des classifieurs naïfs bayésiens et des algorithmes génétiques. Il peut s'agir dans ce cas de l'expression des thématiques très différentes que permettent de traiter ces deux algorithmes. Plus précisément, comme nous l'avons vu, si le naïf bayésien est fortement présent dans le traitement de l'image et l'analyse textuelle, ces domaines sont presque inexistants dans les usages des algorithmes génétiques. Parmi les autres liens négatifs notables, on trouve le couple entre les forêts aléatoires et les naïfs bayésiens, et entre les SVM et les algorithmes génétiques. Là aussi des logiques thématiques semblent expliquer ces oppositions. De la même manière, on trouve aussi des logiques thématiques qui semblent expliquer les attraction positives (rouges) représentées par cette matrice, par exemple entre réseaux de neurones et algorithmes génétiques, ou entre naif bayésiens et SVM.

La logique thématique semble ne plus avoir d'effet lorsqu'on considère la matrice représentant le site *cross-validated*. Les surreprésentations qui apparaissent semblent d'avantage exprimer les rapprochements entre mêmes traditions statistiques. Ainsi la cellule la plus intense de cette matrice est celle entre arbres de décision et forêts aléatoires, on retrouve le rapprochement "naturel" entre les approches *Naive Bayes* et *Bayes Net* qui était négatif dans le cas de *stackoverflow*.

Il semble donc que suivant que les utilisateurs de *stackexchange* identifient leurs questions sur l'apprentissage à un usage de statistique, représenté par *cross-validated*, ou une problématique d'ingénierie, représentée par *stackoverflow*, il existe des parcours d'usages des algorithmes sensiblement différents. En effet, si les compétences statistiques semblent permettre aux utilisateurs de plus facilement se déplacer d'un algorithme à l'autre, ils restent fidèles aux traditions statistiques qui lient les différentes procédures. A contrario, les approches ingénieriques plus pragmatiques, semblent largement guidées par les types de question et de données traités par chaque famille d'algorithme.

Afin d'approfondir les usages dans un contexte purement applicatif, nous analysons, dans la section suivante, l'usage qui est fait de ces algorithmes sur le site de compétitions Kaggle.

4.2 KAGGLE

Si le site Kaggle offre un corpus de données beaucoup plus restreint et moins structuré que ceux envisagés jusqu'à présent, il nous permet d'observer un usage caractéristique de la pratique contemporaine de l'apprentissage : les compétitions de modèles prédictifs et de classification. Après avoir présenté cette plate-forme et les conditions dans lesquelles se déroulent les compétitions (§4.2.1), nous pourrons observer quels algorithmes sont les plus sollicités (§4.2.2) ainsi que leur profil de coprésence (§4.2.3)

4.2.1 *Présentation de Kaggle*

Le site *Kaggle*[4] est une plate-forme de compétitions d'apprentissage artificiel. Fondée en 2010 par Anthony Goldbloom dans la *Silicon Valley*, son principe vise à profiter du caractère générique et immatériel des modèles prédictifs et de classification pour les mettre en concurrence en fonction de critères d'évaluation généralement très précis. Ainsi, une première observation à faire sur les modèles hébergés sur ce site est que l'impératif d'interprétabilité est abandonné et ne fait pas partie des critères de classement des modèles. Pour chaque compétition, on retrouve plutôt une fonction d'évaluation précise qui mesure le succès d'un modèle par un résultat unique. La Figure 34 permet d'observer un exemple des termes d'évaluation, dans ce cas pour la compétition *San Fransisco Crime Classification* qui invite ses participants à trouver le meilleur modèle pour prédire le type de crime perpétré en fonction de sa date et de son lieu, à partir de 12 ans de données.

Chaque compétition possède donc une page équivalente pour écarter toute ambiguïté ou subjectivité dans l'appréciation des modèles. Cette mesure objective permet de classer les candidats avec un seul critère et ainsi de désigner un gagnant unique. La Figure 35 permet d'observer un tel tableau pour la classification des crimes.

Les données déposées sur la plate-forme proviennent de sources très diverses : entreprises, laboratoires de recherche, institutions publiques. Celles-ci recouvrent des domaines d'applications très variés comme la finance, le management, l'aéronautique, la santé publique, la médecine, l'hôtellerie, l'astrophysique, l'assurance, etc. En avril 2016, dernière date à laquelle ont été extrait les données du site pour cette étude, on compte 210 compétitions, dont 12 actives. En effet, la possibilité de participer à une compétition est souvent limitée dans le temps afin de permettre de juger à un moment donné d'un candidat

4. https://www.kaggle.com/

Evaluation

Submissions are evaluated using the multi-class logarithmic loss. Each incident has been labeled with one true class. For each incident, you must submit a set of predicted probabilities (one for every class). The formula is then,

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of cases in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j.

The submitted probabilities for a given incident are not required to sum to one because they are rescaled prior to being scored (each row is divided by the row sum). In order to avoid the extremes of the log function, predicted probabilities are replaced with $max(min(p, 1 - 10^{-15}), 10^{-15})$.

FIGURE 34 – Extrait du site Kaggle[5] détaillant les termes de l'évaluation des modèles pour la compétition *San Fransisco Crime Classification*

Dashboard | Public Leaderboard - San Francisco Crime Classification

This leaderboard is calculated on all of the test data. See someone using multiple accounts? Let us know.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	voltron1985 *	1.95936	17	Tue, 26 Apr 2016 17:45:11
2	–	mehran	2.05079	97	Mon, 11 Jan 2016 15:42:13
3	–	Jghjgfgh	2.06702	12	Sat, 05 Dec 2015 01:44:02
4	–	papadopc	2.11607	111	Tue, 08 Dec 2015 04:54:53 (-32.4d)
5	–	raytrace	2.11638	18	Sat, 05 Sep 2015 19:08:49
6	–	Crime Syndicate	2.13695	20	Sun, 06 Dec 2015 00:04:29
7	–	Wahlen	2.14682	11	Thu, 15 Oct 2015 12:09:40

FIGURE 35 – Extrait du site Kaggle[6] le classement final des participants pour la compétition *San Fransisco Crime Classification*

gagnant qui se voit remettre un prix comme une gratification symbolique, une offre d'emploi, de l'argent.

Parmi les nombreuses fonctionnalités offertes par le site, les utilisateurs peuvent héberger des scripts, c'est à dire des programmes informatiques, associés à une compétition à laquelle l'utilisateur en question est inscrit. Ainsi la plupart de ces scripts visent à relever le défi d'une compétition en utilisant un des langages de programmation compatibles (Python, R ou Julia) et une ou plusieurs des nombreuses librairies d'analyse de données disponibles. Une librairie, ou module, est un ensemble de codes informatiques qui permet à ses utilisateurs de l'invoquer et de réutiliser les fonctions qu'il offre pour ne pas avoir à les implémenter soi-même. Par exemple, dans le language Python, et dans le contexte de l'apprentissage automatique, les librairies les plus utilisées sont en général Scikit-Learn et Thenao.

Des 9358 scripts disponibles sur Kaggle au moment de cette étude, environ la moitié sont en Python (4775) et cette étude ne considère que ceux-là. Une des raisons principales pour n'avoir gardé que les scripts en langage Python est que la procédure pour extraire les modules de chaque script parfois cryptiques et interpréter les relations entre fonction, sous-modules, modules et imports était grandement facilitée par ma connaissance de ce langage. Parmi tous les modules chargés dans les scripts, nous avons sélectionné ceux qui étaient pertinents du point de vue de l'usage des algorithmes d'apprentissage, par exemple SVM ou naïfs bayésiens. Ces modules ont été sélectionnés manuellement et ensuite regroupés dans un dictionnaire pour les lier à leur algorithme de référence. Le dictionnaire est disponible en annexe de cette thèse (§B.2). Ainsi, 1,480 scripts sont représentés ici par le ou les méthodes qu'ils utilisent, et peuvent être liés à la compétition pour laquelle ils ont été créés, et aux utilisateurs qui en sont les auteurs.

4.2.2 *Algorithmes et compétitions*

Le fait d'avoir lié chaque script à un algorithme utilisé dans la procédure d'apprentissage permet d'observer dans un premier temps simplement le nombre d'occurrences de chaque famille d'algorithme, comme représenté par la Figure 36. Ainsi, on peut voir que deux des familles observés dans le champ académique ne sont pas du tout présentes : les algorithmes génétiques et les réseaux bayésiens. Comme nous l'avons vu, ces deux algorithmes figurent parmi ceux qui nécessitent beaucoup d'adaptations et de paramètres vis-à-vis du contexte et des données auxquelles ils sont appliqués. Cependant, cela ne suffit pas à expliquer leur absence de Kaggle étant donné que les réseaux de neurones ont cette même caractéristique et sont toutefois particu-

lièrement sollicités par les utilisateurs. Une raison à cette différence peut-être que les réseaux de neurones sont à l'honneur dans le littérature académique et attirent depuis quelques années bon nombre des espoirs de l'IA et incitent ainsi à plus d'efforts de leurs utilisateurs qui peuvent profiter de nombreux matériaux pédagogiques et de vulgarisation qui accompagnent cet enthousiasme. Peut-être aussi les réseaux de neurones sont-ils tout simplement plus efficaces. De manière plus concrète, on voit dans l'annexe §B.2 que les librairies de programmation sollicitées pour les réseaux de neurones sont le fruit d'efforts très récents : Thenao et Nolearn ont été créés en 2012, Lasagne et Keras en 2015, et le module correspondant dans la librairie plus généraliste Scikit-learn n'est disponible depuis 2016 qu'en version de développement. Ainsi l'ensemble des efforts des participants utilisant les réseaux de neurones repose sur des outils seulement créés depuis quelques années et sont donc le fruit de l'investissement massif d'une communauté dans l'abstraction et la simplification des procédures connexionnistes. Si les réseaux bayésiens et les algorithmes génétiques continuent, comme nous l'avons vu dans le chapitre précédent, à être des pistes de recherche actives dans leurs améliorations et dans leurs applications, ces méthodes ne sont pas à l'honneur des récents développements de l'apprentissage et, en partie pour cette raison, n'attirent pas l'effort d'une communauté à rendre leurs outils plus accessibles et génériques.

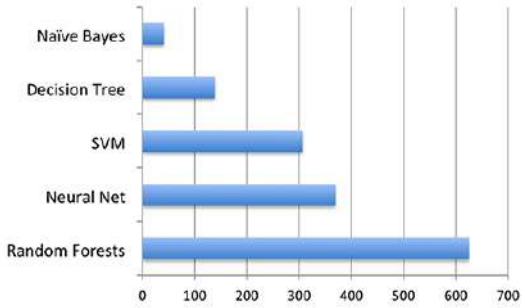


FIGURE 36 – Nombre de scripts python utilisant chaque algorithme sur Kaggle.

Les trois algorithmes qui apparaissent comme les plus fréquents dans la Figure 36 sont les forêts aléatoires, les réseaux de neurones et les SVM. Tous les trois font partie des familles d'algorithmes auxquelles s'identifie la "nouvelle culture de modélisation" formulée par Breiman [20]. Étant donné que la plupart des compétitions hébergées sur Kaggle classent leurs participants selon un score unique de précision de la tache de classification ou de prédiction envisagée, il n'est pas étonnant de trouver davantage sollicités sur la plate-forme ceux qui privilégient l'efficacité par rapport à d'autres qualités possibles du modèle, comme leur interprétabilité. Kaggle semble donc, par ce simple mode de classement, défavoriser à la fois les algorithmes peu

sollicités dans les derniers développements en apprentissage artificiel mais encore priviligier ceux qui incarnent le mieux l'ambition et les contraintes de cette nouvelle culture de la modélisation.

En détaillant la place des algorithmes par compétition, on obtient la Figure 37 qui permet d'observer, pour chaque compétition pour lesquelles nous avons pu identifier plus de 10 algorithmes utilisés, le pourcentage de chacun dans les solutions proposées par les participants.

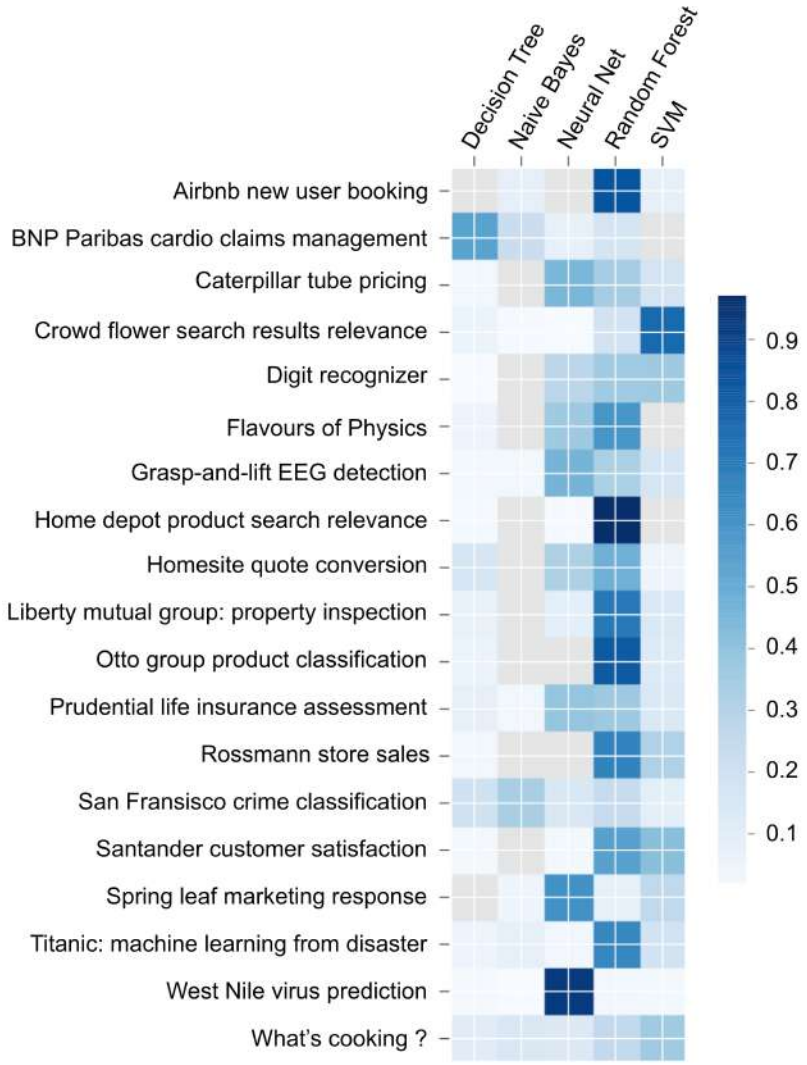


FIGURE 37 – Place de chaque algorithme dans les contributions aux compétitions sur Kaggle.

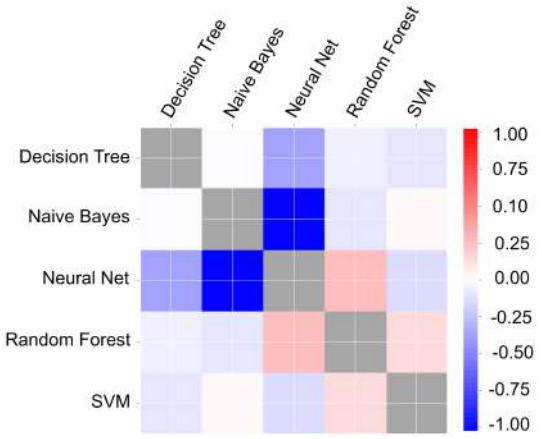
De manière plus générale, on voit que la plupart des compétitions sont l'objet de procédures d'apprentissage issues d'un seul algorithme principal. Aussi, une minorité de compétitions est ouverte à une plus grande variété d'algorithmes, laissant alors penser qu'une procédure en particulier ne fait pas une différence majeure avec les autres. Dans les cas où un seul algorithme domine, la majeure partie de ces com-

pétitions rassemblent principalement l'utilisation de forêts aléatoires, montrant ainsi que la forte présence de cet algorithme sur toute la plate-forme et répartie de manière relativement uniforme dans une majorité de compétitions. On retrouve néanmoins quelques exceptions, notamment pour les réseaux de neurones dans les compétitions *West Nile Virus prediction* et *Springleaf marketing response*.

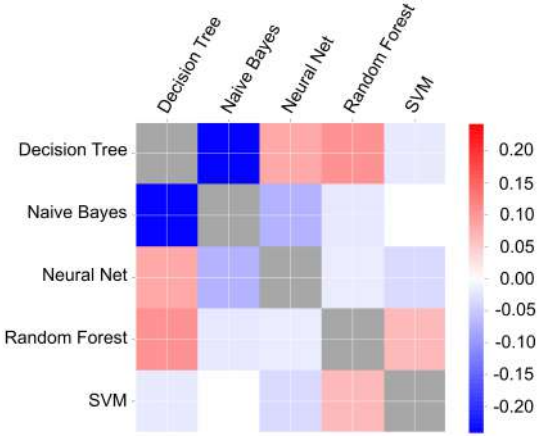
4.2.3 *Co-présences des algorithmes*

En ayant lié tous les participants à Kaggle avec les algorithmes qu'ils ont sollicités lors de leurs multiples soumissions aux compétitions, on peut alors observer la coprésence des algorithmes par participant en suivant le même méthode que celle décrite dans la section §3.3.3. On trouve 614 participants ayant publié au moins un script python dont on a pu identifier les imports. Parmi ceux-là, la Figure 38a nous permet d'observer les coprésences d'algorithmes pour les 208 utilisateurs ayant sollicité plus d'un type d'algorithme. On considère qu'un participant sollicite plusieurs algorithmes en confondant tous les imports de tous les scripts dont il est l'auteur ou qu'il a utilisé, et en comptant le nombre d'algorithmes qui apparaissent. En suivant cette méthode, la Figure 38b permet d'observer les coprésences d'algorithmes pour les 29 compétitions pour lesquelles nous possédons des scripts dont les imports sont identifiés. Parmi celles-ci, seulement 24 ont fait l'objet de procédures d'apprentissage d'au moins deux familles d'algorithmes et se trouvent donc représentées par la matrice.

La coprésence par participant indique qu'en plus d'être les algorithmes les plus fréquents, les forêts aléatoires, les réseaux de neurones et les SVM ont tendance à se retrouver ensemble dans les soumissions d'un participant. Cela peut indiquer que la forte présence de ces algorithmes sur l'ensemble de la plate-forme Kaggle, en plus d'être répartie de manière homogène entre compétitions, est aussi l'œuvre d'un bon nombre d'utilisateurs qui ont tendance à explorer des solutions alternatives au sein de cette ensemble d'algorithmes. Il semble que le paysage soit différent quand il s'agit de la coprésence des compétitions par algorithme. En effet, les réseaux de neurones et les forêts aléatoires apparaissent comme surreprésentées avec les arbres de décisions. Une des raisons pour expliquer cela, est que les arbres de décisions sont utilisés en amont des procédures plus performantes afin de réaliser une première exploration des bases de données des compétitions. Dans ce cas, on pourrait dire qu'il s'agit de méthodes désuètes d'apprentissage qui subsistent dans cet écosystème technique parce que leur intelligibilité soutient les premiers pas d'exploration des données.



(a) Coprésence par participant



(b) Coprésence par compétition

FIGURE 38 – Matrice de coprésence des utilisateurs (*a*) et des compétitions (*b*) par algorithme.

RÉSUMÉ DU CHAPITRE 4

Dans ce chapitre, nous nous sommes intéressés à des usages plus récents de l'apprentissage artificiel rendu plus accessible par de nombreux outils, comme des librairies de programmation, et les contenus pédagogiques qui les accompagnent. Ainsi, on s'est concentré sur les traces disponibles des utilisateurs de plusieurs forums de questions-réponses (*Stackexchange*) et d'une plateforme de compétitions de modèles prédictifs et de classification (*Kaggle*).

Pour les forums, la simple identification de ceux étant pertinents à notre étude nous a permis de retrouver les domaines envisagés dans le contexte académique, lors du chapitre précédent. Les réseaux de co-occurrences des mots-clés les plus associés à l'apprentissage artificiel nous ont permis de regrouper les problématiques les plus traitées, comme les langages, méthodes de calcul, contraintes du type de données, etc. Ces problématiques nous ont permis d'apporter plusieurs explications à la co-présence des algorithmes dans les usages des utilisateurs étudiée par la suite.

L'analyse des traces des compétitions sur *Kaggle* révèle que le mode d'évaluation uniforme privilégié par la plateforme favorise plusieurs algorithmes au détriment d'autres, ou de leur absence. La co-présence des algorithmes, autant pour les utilisateurs que pour les compétitions, semble aussi fortement déterminée par les contraintes de performance induites par la plate-forme.

CONCLUSION

A travers l'examen de plusieurs procédures d'apprentissage artificiel nous avons pu voir que la référence à l'apprentissage ne relève pas d'un axiome scientifique unique, mais de l'entremêlement de plusieurs traditions scientifiques : statistique, mathématique, et plus récemment informatique. Si l'on a rendu raisonnablement compte de ce champ de recherche en évoquant l'*apprentissage par l'expérience*, ou l'*automatisation par induction*, il semble également que des dynamiques internes fortes s'ancrent dans l'histoire d'hypothèses vieilles de plusieurs siècles (bayésianisme, régression, évolution, etc.) et de leurs parcours relativement autonomes vers une certaine forme d'IA. Au niveau le plus général, l'apprentissage artificiel est une solution statistique à des approches quantitatives et informatiques de résolution de problèmes. Plus récemment, on assiste à l'émergence du mouvement des *Data Science* dont le dénominateur commun, l'analyse de données, semble progressivement subsumer l'apprentissage.

La disponibilité croissante des données et des capacités de calculs, et les contraintes juridiques et opportunités économiques qui entourent leurs usages et leurs mises en relations, sont probablement les déterminants les plus forts de l'histoire contemporaine du *machine learning*. Ainsi, en s'intéressant dans cette thèse à ses algorithmes et non à son économie, on a fait le choix de rendre compte des différentes *épistémès* qui rythment la course vers une procédure d'apprentissage suffisamment universelle pour qu'on n'ait plus à en rendre compte. Dans une fenêtre de temps de quelques décennies, on a saisi une poignée d'algorithmes qui, comme on l'a vu, sont les centres de diverses *doxas*, hypothèses scientifiques et communautés de recherche qui se structurent autour d'un ensemble de contraintes tissées par les types de données, les méthodes de calcul et les compromis d'usage et de performance. On a ainsi pu observer l'évolution récente de ce champ de recherche et comment chacune de ses "tribus" fait appel à certaines métaphores pour imiter l'apprentissage, d'éléments de théorie scientifique et d'implémentation algorithmique propre.

Afin de pouvoir observer la structuration et les dynamiques d'interaction de ces "tribus", on s'est intéressé ensuite à l'activité académique qu'elles suscitaient en prenant appui sur de larges corpus de données issues de *Web of Science* pour chaque algorithme considéré. Les activités ainsi perçues montrent que, dans le contexte général d'une forte croissance de la publication scientifique, l'apprentissage artificiel est un champ de recherche qui accompagne ce mouvement, tant par l'augmentation de son nombre d'auteurs, que par les origines

géographiques de leurs contributions. Fort de ces constatations, on a pu établir une méthode d'analyse des réseaux de co-citations pour chaque corpus et ainsi détecter les communautés thématiques qui s'y expriment le plus fréquemment. Ainsi, a-t-on pu montrer que les thématiques observées dans le plus grand nombre de corpus (biologie, chimie) entretiennent très peu de liens avec la recherche théorique sur l'algorithme lui-même. Aussi, bon nombre de thématiques n'apparaissent-elles que dans un seul corpus mettant en lumière à quel point les contraintes de calcul, l'interprétabilité des modèles et les traditions scientifiques peuvent être déterminantes pour le choix d'une procédure d'apprentissage. Enfin, on a pu rassembler quelques éléments d'identification d'un *opportunisme méthodologique* du *machine learning* : il semble en effet qu'il soit plus simple pour un auteur de se déplacer d'une thématique à une autre que d'une méthode d'apprentissage à une autre.

En dernier lieu, nous nous sommes intéressés à des pratiques plus contemporaines et généralisées de l'apprentissage artificiel au travers de sites de questions-réponses de développeurs (*Stackexchange*) et d'une plateforme de compétitions de modèles prédictifs et de classification (*Kaggle*). Pour les services de questions-réponses, la simple identification des sites les plus pertinents nous a permis de retrouver certains résultats obtenus sur l'analyse des corpus académiques tels que la présence des mêmes domaines d'application spécifiquement liés à chaque type d'algorithme. Les réseaux de co-occurrences des mots-clés référençant les questions permettent de regrouper les problématiques privilégiées (langage, thématique, méthodes de calcul, etc.) par chacune de ces thématiques, permettant d'expliquer en partie la co-présence des algorithmes dans le parcours des utilisateurs. Dans le cas de la plate-forme de compétitions de modèles, on a pu observer comment le recours à un seul et unique mode d'évaluation tend à homogénéiser les stratégies des participants qui privilégient alors un ou deux algorithmes en particulier (forêts aléatoires, réseaux de neurones). La co-présence des algorithmes, autant pour les utilisateurs que pour les compétitions, semble aussi fortement déterminée par les contraintes de performance induites par la plate-forme.

De manière plus générale, cette étude de l'apprentissage artificiel par le prisme de ses algorithmes nous a permis de cartographier les communautés d'auteurs, de contraintes, de thématiques et références qui se structurent autour de ceux-ci. Pourtant, si les comportements saisis par ces analyses sont révélateurs des usages contemporains du *machine learning*, l'extrême actualité des débats sur ces méthodes semble se concentrer sur leurs utilisations dans le cadre de la prise de décisions publiques. En effet, à mesure que les techniques d'apprentissage s'étendent à un nombre croissant de domaines d'application, une partie du débat autour de ces pratiques se cristallise sur leurs rôle

en tant qu'artefact socio-technique. Autrement dit, chacun des "styles de pensée" des algorithmes d'apprentissage revêt un positionnement politique, économique et social particulier

Pour donner un cadre à ce débat, les acteurs font souvent appel à des représentations caricaturales de la réalité que recouvre l'apprentissage artificiel, autant du point de vue des données disponibles que de leurs usages. On retrouve ainsi souvent le concept de *société prédictive* pour désigner le schéma dans lequel l'apprentissage serait capable de déceler n'importe quel type de corrélation, où toutes les données sont conservées, structurées et potentiellement mises en relation.

Mais si la société prédictive ne reflête pas la situation actuelle, elle pointe bien les ambitions des promoteurs les plus forcenés de l'IA. À titre d'exemple, Kevin Kelly, "prophète du silicone" [44] dans la vallée éponyme, annonce dans un article de prospective sur l'IA ayant rencontré beaucoup de succès : "En fait, les *business plans* des prochaines 10,000 start-ups sont faciles à prédire : Prendre *X* et l'augmenter d'IA." [7] [57]. Dans le même sens, on retrouve souvent dans les médias nombre de projets ayant l'ambition d'intégrer de l'IA dans la médecine, les transports ou la réalité virtuelle. La frontière entre avoir raison trop tôt ou avoir simplement tort est ténue, mais il n'est pas forcément nécessaire que les éléments d'une *société prédictive* soient imminents pour que son imaginaire soit révélateur de notre époque.

Dans le cadre de la décision publique, que l'on définira simplement comme s'appliquant à un domaine jugé d'intêrét général dans un groupe considéré, un concept régulièrement brandi aussi bien par les institutions, les chercheurs ou les simples citoyens sont ceux de la justice sociale (*fairness*) et de la transparence qu'implique l'usage d'algorithmes. Ces thématiques font l'objet de plusieurs rapports depuis 2013, notamment aux EUA de la *National Acedemy of Sciences* [25] et de la Maison Blanche [50, 51]. À partir de 2014, on voit se former l'atelier *Justice, transparence et responsabilité dans l'apprentissage artificiel* [8] dans plusieurs des plus importantes conférences académiques sur le *machine learning*, notamment ICML et NIPS.

Alors que la problématique de la transparence rejoint celle de l'interprétabilité des modèles que nous avons abordée à de nombreuses reprise dans cette thèse, celle de "justice" (*fairness*) est beaucoup plus spécifique à la nature socio-politique de l'apprentissage. En ce sens, la notion de justice fait l'objet de plus d'attention de la part de la presse écrite - notamment en 2015 dans le *New York Times* [77] et *The Atlantic* [59] - que celle de transparence. Un groupe de recherche s'est aussi dédié à cette question depuis peu au sein de l'Université de

7. "In fact, the business plans of the next 10,000 startups are easy to forecast : Take X and add AI."

8. "Fairness, Accountability, and Transparency in Machine Learning", http://www.fatml.org/

Haverford [9], et en mai 2016 IEEE organise une conférence internationale dédiée à cette question [10]. La question est la suivante : est-ce que l'apprentissage artificiel porte en son sein une forme de justice ou d'injustice structurellement liée à ses méthodes ? Si oui, quelles sont ses caractéristiques ?

Deux des concepts souvent sollicités pour aborder la question de la "justice" d'un algorithme d'apprentissage sont ceux de *discrimination* et de *biais*. On peut noter que l'émergence de ces deux concepts pour discuter de l'identité politique des algorithmes est peut-être due au fait qu'ils ont tous deux des sens importants, bien que différents, dans l'analyse sociale et dans celle mathématique des algorithmes. En effet, d'un point de vue politique, la discrimination désigne l'action de distinguer de manière juste ou injuste, légitime ou illégitime un groupe social ou un individu, d'un autre. D'un point de vue purement mathématique, la discrimination est un synonyme de *distinction*, et, à ce titre, on retrouve plusieurs algorithmes qui y font référence, comme l'*Analyse Discriminante Linéaire*. Les sens du mot *biais* sont plus proches et désignent tous deux une erreur systématique ou une simplification. Néanmoins son usage politique fait beaucoup plus référence à son étymologie, du grec, "violence", alors que son sens mathématique formalise l'idée de simplification, et l'oppose à la *variance*.

Parmi les exemples souvent cités pour débattre de la nature discriminante des algorithmes, le cas mis en lumière par les travaux de Datta et al. [28] est souvent pris en exemple. Les auteurs de cette expérience ont réussi à mettre à jour un biais révélateur de l'algorithme gérant l'affichage de publicités du moteur de recherche Google (*Google AdWords* [11]). Il a ainsi été mis en évidence que les femmes se voient proposer des offres d'emploi en moyenne bien moins prestigieuses ou moins bien payées que celles proposées aux hommes. Fort des multiples descriptions de procédures d'apprentissage rencontrées pendant cette thèse, on peut spéculer comme suit sur les origines de ce biais discriminant systématique dans le système mis en place par Google.

Google fournit un outil de recherche d'information sur le web. Les utilisateurs peuvent être la plupart du temps identifiés grâce à leur compte ou des techniques de traçage (cookies, adresse IP). Ainsi, Google peut associer à une recherche un profil et donc les recherches passées, ainsi que d'autres variables, parfois directement renseignées par l'utilisateur, soit inférées ou prédites. Parmi ces variables, le genre de l'utilisateur est un moyen d'appuyer une distinction à partir de la-

9. http://fairness.haverford.edu/index.html
10. IEEE PDDM 2016 : International Workshop on Privacy and Discrimination in Data Mining : http://pddm16.eurecat.org/
11. https://www.google.fr/adwords/

quelle s'exprime une corrélation entre une variable et une attente de certains résultats plutôt que d'autres. C'est donc en grande partie par le comportement passé des utilisateurs et leurs caractéristiques que le moteur de recherche identifie les résultats pertinents. Certaines de ces variables sont peu ou pas polémiques, par exemple lorsque un habitant de Toulouse fait la requête "cinéma" il se voit proposer les cinémas de sa ville et non de Paris. Pourtant, si on prend en compte toutes les requêtes pour "cinéma" en France, il est probable que le plus grand nombre viennent de Paris. Mais la variable de localisation *distingue* la pertinence de renvoyer les résultats toulousains plutot que parisiens. D'autres variables sont considérées comme sensibles et génèrent beaucoup de polémiques, comme le genre ou l'origine ethnique. S'appuyer sur cette variable pour raffiner le contenu est considéré comme un acte discriminatoire. Néanmoins, le simple fait de s'appuyer sur ce type de variable n'implique pas nécessairement que la prédiction soit de nature discriminatoire. Par exemple, on peut imaginer qu'une prédiction pour une requête de shampoing ne serait pas considérée comme discriminatoire si elle parvenait à faire ressortir des produits spécifiques aux particularités capillaires d'utilisateurs d'un genre donné (homme, femme, etc) ou même d'une couleur de peau donnée (blanc, noir, etc). C'est donc bien le croisement de variable dites "sensibles" et d'un domaine d'ordre "politique" qui confère à une prédiction son caractère discriminatoire, comme c'est le cas, dans notre exemple, pour le genre et un entretien d'embauche.

On peut définir une variable dite "sensible" comme l'ensemble des marqueurs d'inégalité structurelle et souvent ancienne d'une société. Par exemple dans le cas de l'ethnie, cela renvoie souvent au passé colonial ou esclavagiste d'une nation qui s'exprime encore dans l'accès aux richesses, moyens de productions, représentation politique, etc. Dans le cas du genre, il s'agit là aussi de l'expression d'une situation de domination très ancienne qui explique encore des inégalités de salaires, des rôles sociaux et politiques distincts, etc.

Le caractère "politique" d'un domaine de discrimination peut être défini comme ayant trait à un mécanisme au sein de l'organisation de la société au moment duquel des choix stratégiques doivent être faits. Par exemple, on retrouve dans ce cas, la sélection des individus à différentes étapes de leur vie et qui détermine en partie leur accès à plusieurs types de ressources comme la sélection à l'école, à l'embauche, aux concours, etc.

Lorsque, dans une société prédictive, un algorithme sélectionne la population pour l'accès à une ressource stratégique, il reproduit les motifs qu'il saisit dans les données, la situation que celles-ci permettent d'observer par le passé. Ainsi, il n'est pas la source de la discrimination mais celui qui la reproduit, et non celui qui la produit. Il prédit, et donc il reproduit. Si on reprend l'exemple des offres d'emplois

faites par Google AdWords, les algorithmes qui en sont arrivés à présenter des annonces moins prestigieuses aux utilisateurs féminins ont probablement seulement reproduit le motif connu et maintes fois étudié d'écart de confiance en soi ("confidence gap") qui sépare les genres dans l'appréciation de leurs compétences [56, 92]. Cet écart de confiance en soi est lui-même probablement une expression de la situation dominante (statut, salaire) des hommes qui peut être observée dans nombreux cadres professionnels.

Ainsi, le biais dont on accuse les algorithmes d'apprentissage n'est pas la distance entre le modèle que produisent ces algorithmes et la réalité, mais bien celle entre la réalité qu'ils modélisent et l'intention ou l'axiome politique d'un domaine. L'algorithme devient une expression ubiquitaire de cette "distance", ou inégalité. En matière de genre comme d'ethnie, la *doxa* politique est généralement positionnée sur une égalité de droits acquise il y a plusieurs décennies. Les comportements sont, quant à eux, en partie contraints par la lente évolution d'une société qui porte encore beaucoup de signes et de déterminants des inégalités passées. Pour palier ce décalage systématique, nombre de législations et de politiques publiques, dites de *discrimination positive*, accompagnent depuis la seconde moitié du XX^{e} siècle, les voeux politiques d'égalité des droits. En France, par exemple, on dispose d'aides financières pour faire ses études en fonction du salaire de ses parents, il y a des amménagements d'enseignement en fonction de codes postaux, certains mandats politiques sont soumis à des quotas de genre. Dans le même sens, aux EUA, il existe des quotas ethniques pour l'entrée dans certaines universités ou pour occuper certains postes dans la fonction publique. Il s'agit d'autant de mesures dont une des ambitions est de permettre et d'accélérer l'évolution d'une société vers un idéal politique d'égalité des droits.

Ce qui caractérise un modèle prédictif ou de classification dans le contexte d'une société prédictive qui en fait usage à des étapes stratégiques de sélection, c'est de renforcer les comportements observés et non l'axiome politique. En l'état, les algorithmes d'apprentissage auront tendance à renforcer les contraintes et déterminants des inégalités structurelles d'une société. Ils auront aussi tendance à les amplifier, en ce qu'ils influencent un nombre de comportements croissants et plus intimes (relations sociales, recherche d'informations, etc.). Cependant, en même temps qu'il deviennent de nouveaux vecteurs de discrimination, les dispositifs de prédiction et de classification deviennent aussi des nouveaux terrains possibles de discrimination positives, c'est à dire de nouvelles opportunités pour permettre ou accélerer l'évolution d'une société vers son idéal politique.

Tout comme cette conclusion, la plupart des travaux sur cette problématique récente se contentent souvent de pointer le problème pour mettre en évidence les questions d'"injustice" et l'opportunité de "jus-

tice" que représentent les algorithmes d'apprentissage dans un idéal-type de société prédictive. Néanmoins, il convient de citer ici les travaux très récents de Muhammad Bilal Zafar et al. [12, 13] qui vise justement à inclure dans la procédure d'apprentissage des contraintes de "justice". En distinguant les variables sensibles des autres, la méthode de Zafar propose d'inclure dans la phase d'apprentissage l'objectif d'agir sur la discrimination observée et ainsi de l'amoindrir. Il est intéressant de noter que sa méthode ne s'applique pour l'instant qu'à la régression logistique et aux SVM, alors que d'autres méthodes comparables ne proposaient jusque là des solutions que pour les modèles naïfs bayésiens [21] ou les arbres de décisions [54]. Il semblerait donc que, comme nous l'avons vu, chaque algorithme opére un compromis entre des aspects techniques, par exemple de distribution du calcul, d'interprétabilité et de performance, et que la "justice" puisse devenir à la fois un élément de ce compromis et un determinant fort de l'évolution de ces algorithmes dans des domaines d'application d'ordre plus politique.

ANNEXES

A

RÉSEAUX DE CO-CITATIONS - *WEB OF SCIENCE*

Cette annexe permet au lecteur d'inférer lui même la nature des communautés thématiques décrites et analysées dans le Chapitre 3 (§3.2.2). Il s'agit donc de la reproduction de tous les réseaux représentés dans la Figure 23.

Chaque communauté détectée est décrite par ses 3 mots-clés (**gras**) et ses 3 journaux (*italique*) les plus fréquents.

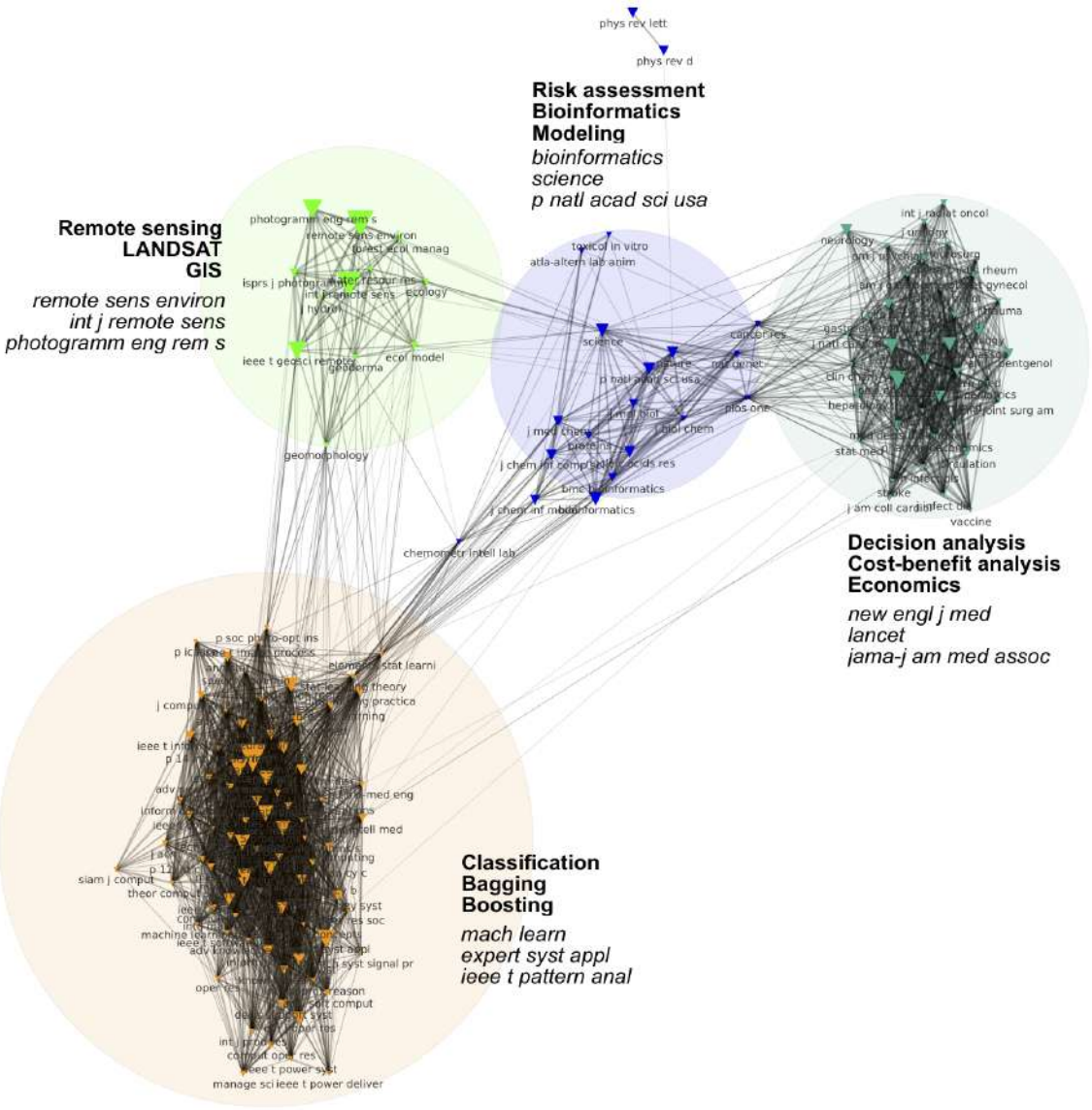


FIGURE 39 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Decision Tree*.

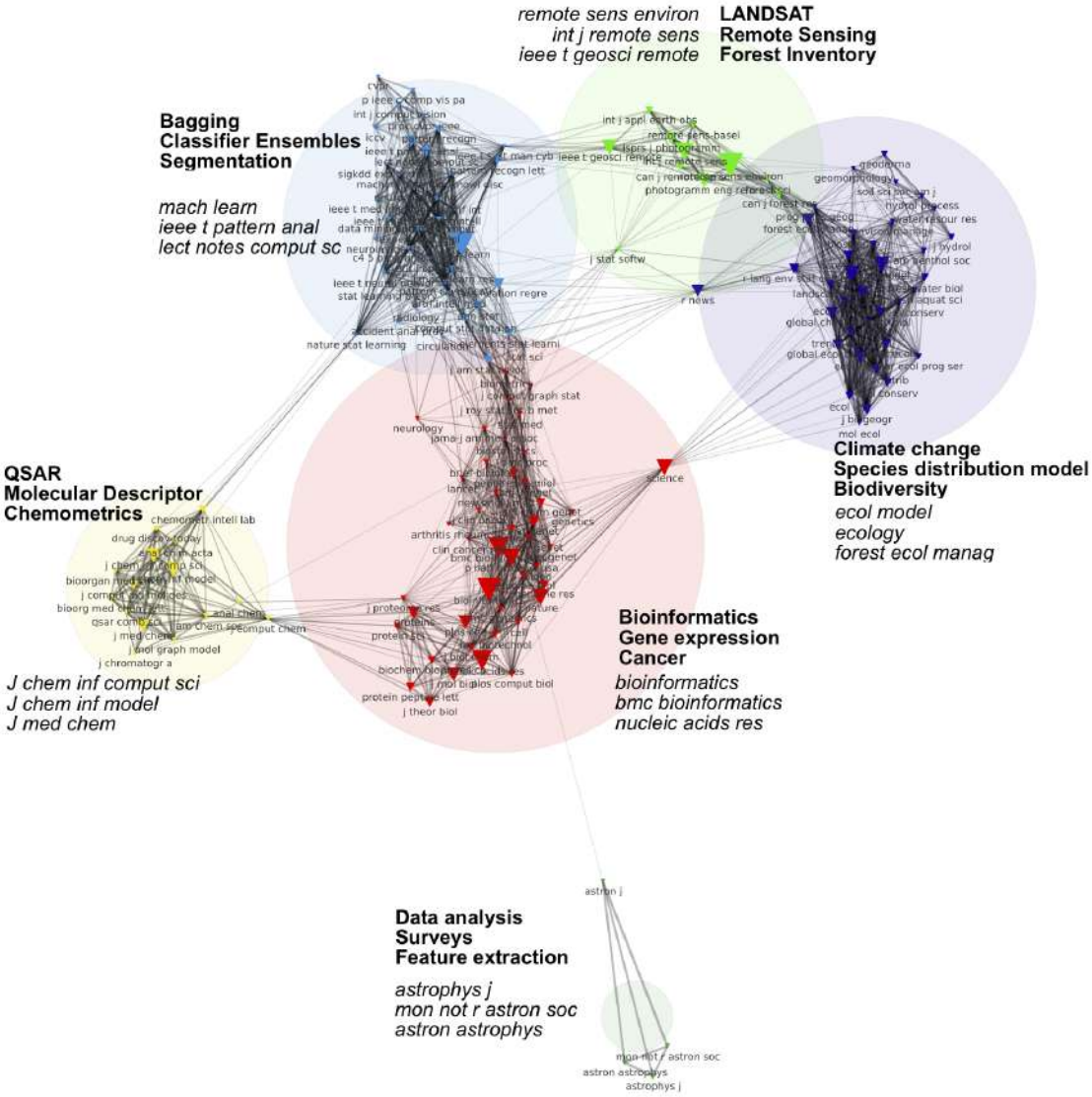


FIGURE 40 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Random Forest*.

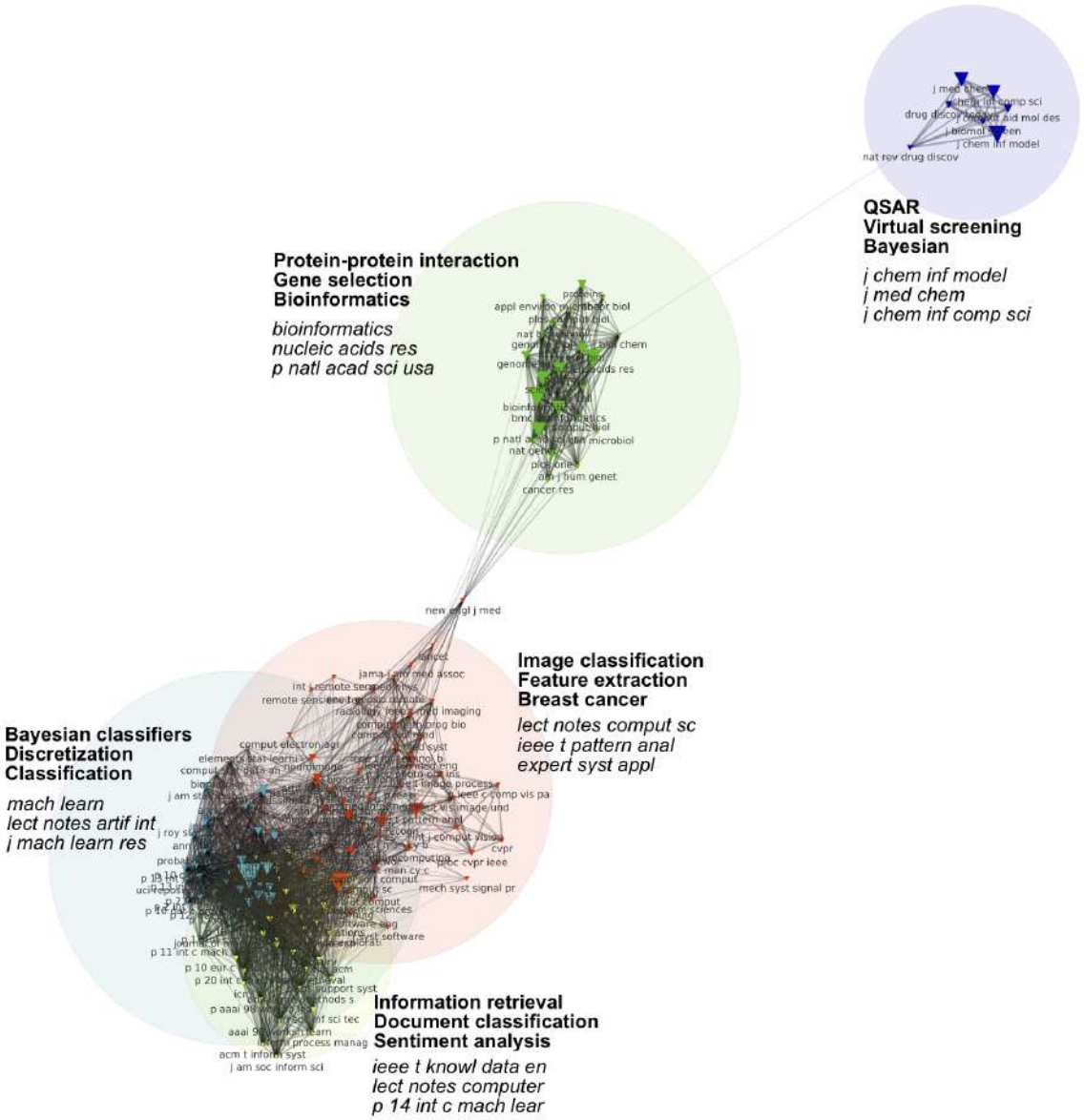


FIGURE 41 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Naive Bayes*.

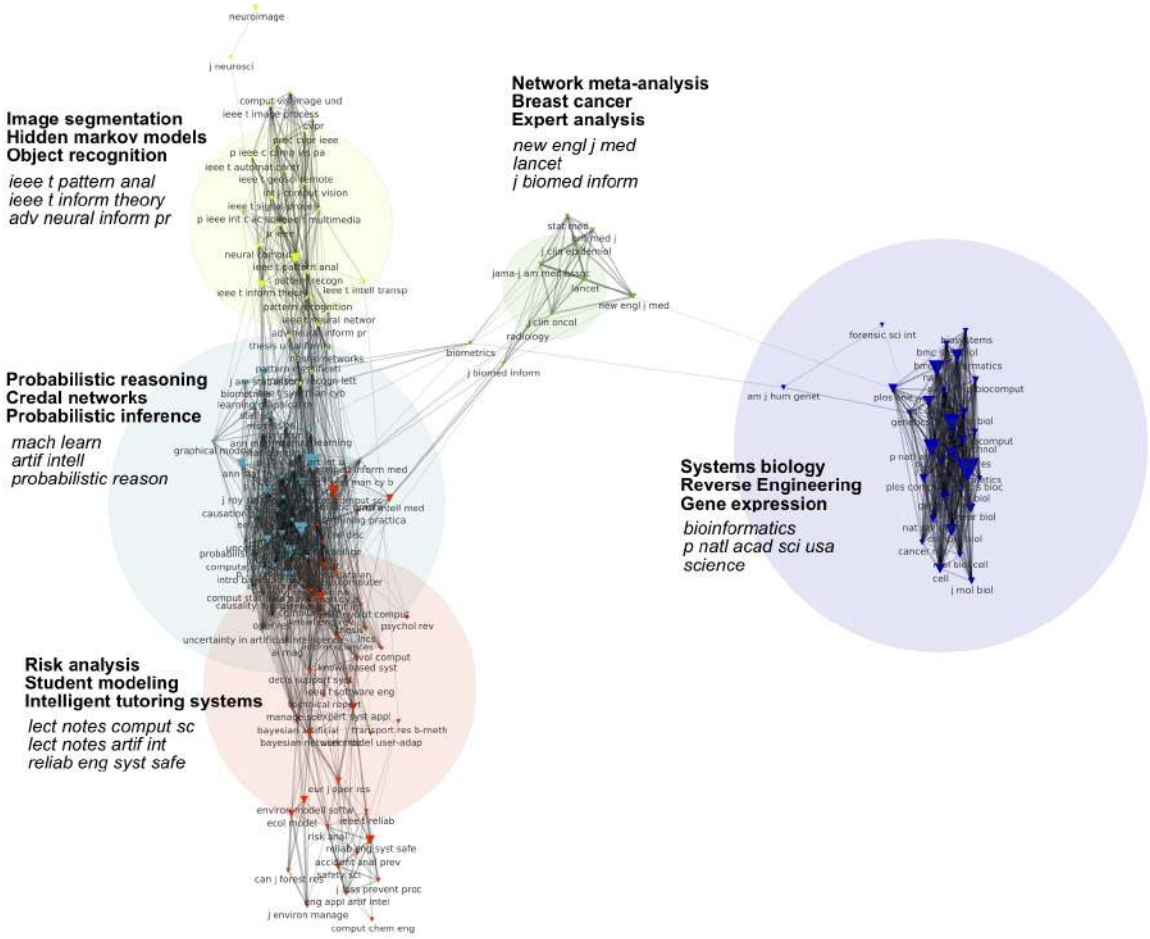


FIGURE 42 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Bayes Net*.

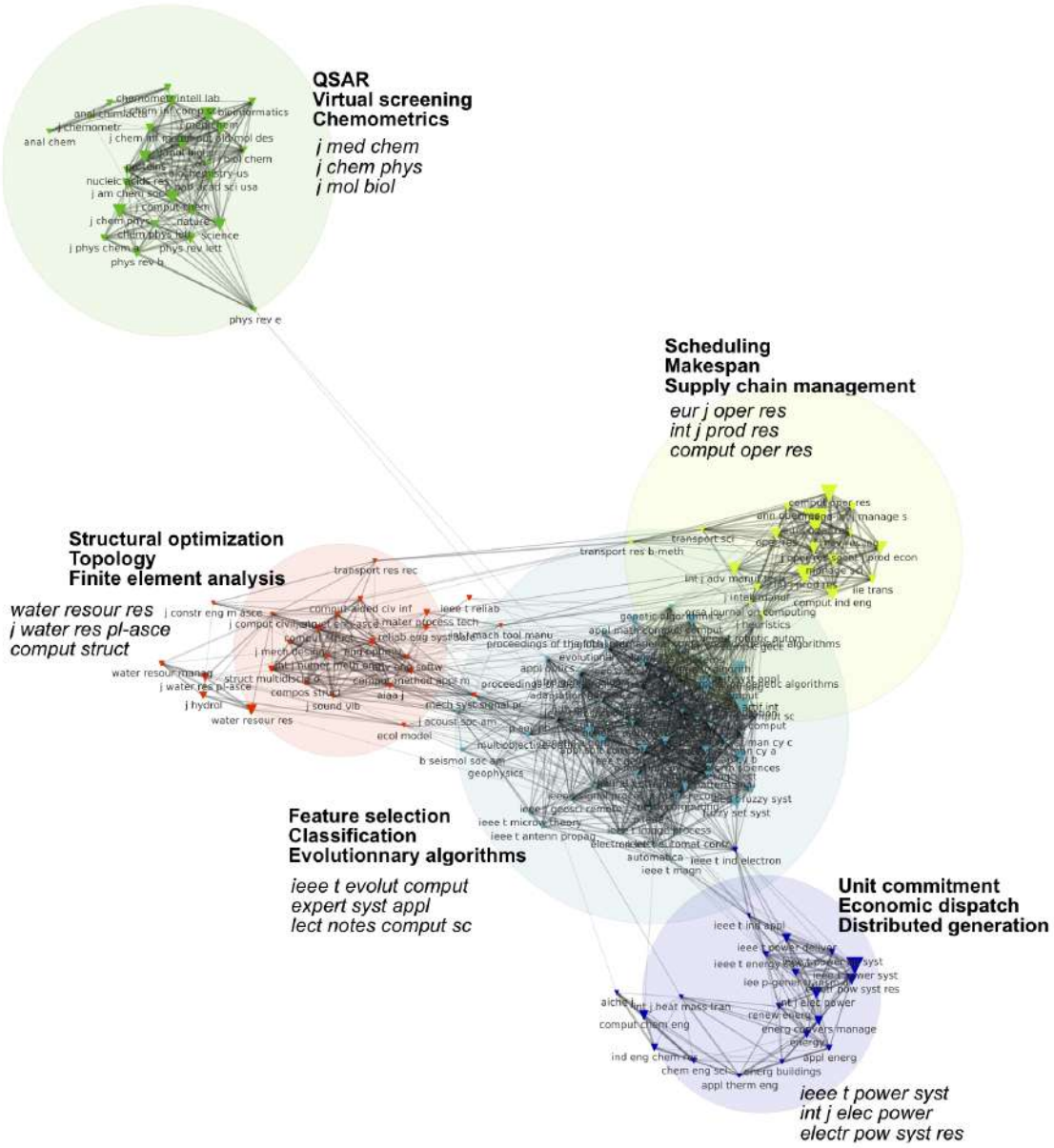


FIGURE 43 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Genetic Algorithm*.

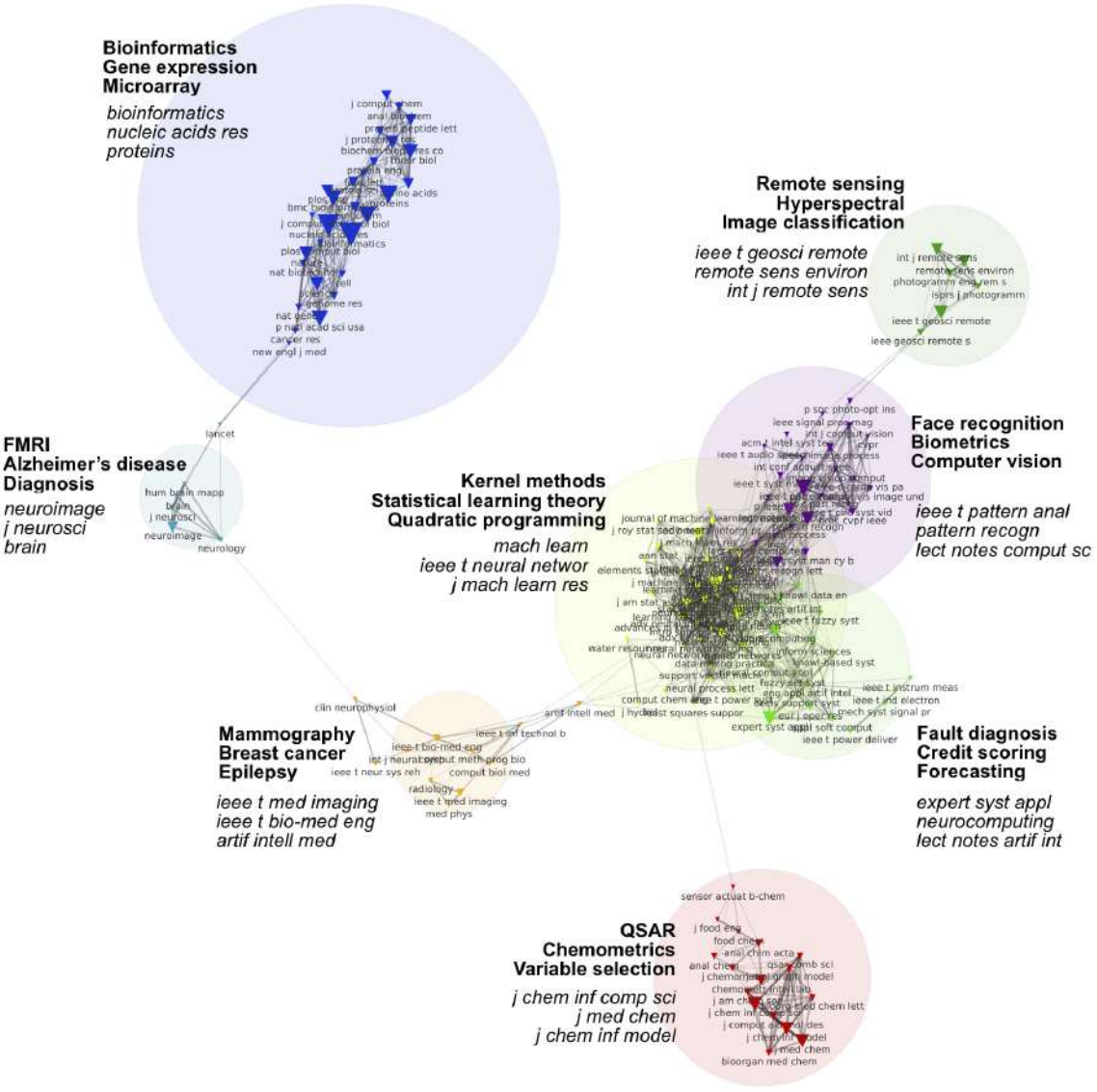


FIGURE 44 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *SVM*.

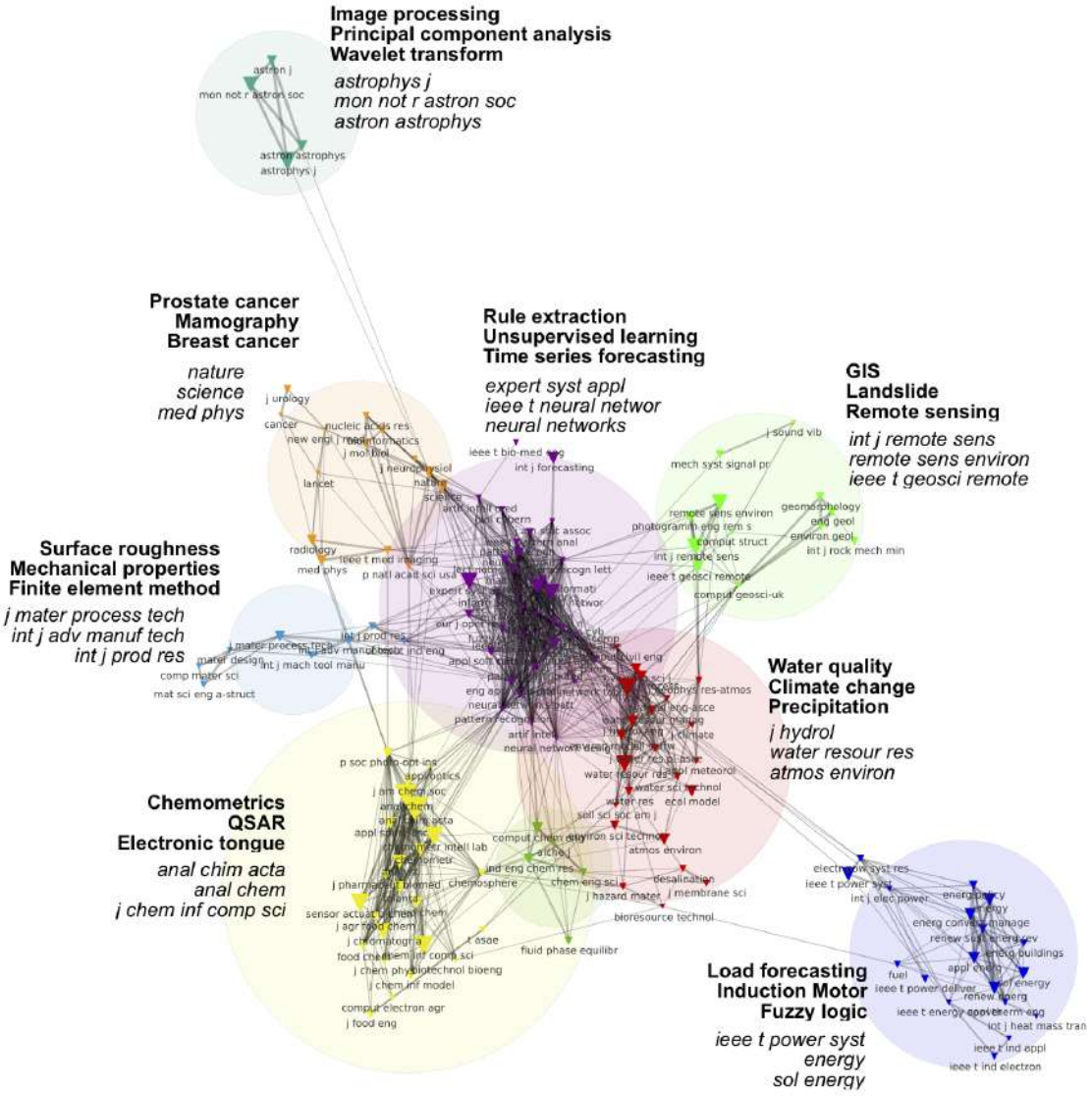


FIGURE 45 – Réseau de co-citation des 150 journaux les plus cités dans le corpus *Neural Net*.

B

DICTIONNAIRES

Cette annexe présente les dictionnaires utilisés pour le chapitre 4 afin d'attribuer dans l'étude des scripts Kaggle et des Questions sur *Stackoverflow* et *Cross-Validated*, à chaque import de code et à chaque mot-clé, l'usage d'un algorithme.

B.1 STACKEXCHANGE

Bribe de code 5 – Dictionnaire de correspondance entre les mots-clés des questions sur *Stackoverflow* et *Cross-Validated* et algorithmes étudiés

```
{'BayesNets': ['bayesian-networks', 'bayes-network', 'bayesian-
    network'],
 'DecisionTree': ['decision-tree', 'cart'],
 'GenAlgo': ['genetic-algorithms',
             'evolutionary-algorithms',
             'evolutionary-algorithm',
             'genetic-algorithm',
             'genetic-programming'],
 'NaiveBayes': ['naive-bayes', 'naivebayes'],
 'NeuralNets': ['neural-network',
                'conv-neural-network',
                'recurrent-neural-network',
                'neural-networks',
                'deep-learning'],
 'RandomForest': ['random-forest'],
 'SVM': ['svm', 'libsvm', 'svmlight', 'libsvmsharp', 'structured-
    svm']}
```

B.2 KAGGLE

Bribe de code 6 – Dictionnaire de correspondance entre les imports de code dans les scripts python sur *Kaggle* et les algorithmes étudiés

```
{'DecisionTree': ['sklearn.tree.DecisionTreeClassifier',
                  'sklearn.tree',
                  'sklearn.tree.DecisionTreeRegressor'],
 'NaiveBayes': ['sklearn.naive_bayes.GaussianNB',
                'sklearn.naive_bayes.BernoulliNB'],
 'NeuralNet': ['keras.models.Sequential',
               'keras.layers.core.Dense',
               'keras.layers.core.Dropout',
               'keras.layers.core.Activation',
               'theano',
               'nolearn.lasagne.NeuralNet',
               'lasagne.updates.nesterov_momentum',
               'keras.utils.np_utils',
               'theano.tensor',
               'lasagne.layers.InputLayer',
               'lasagne.layers.DropoutLayer',
               'lasagne.layers.DenseLayer',
               'lasagne.objectives.binary_crossentropy',
               'keras.layers.advanced_activations.PReLU',
               'theano.tensor.nnet.sigmoid',
               'keras.layers.normalization.BatchNormalization',
               'keras.optimizers.Adadelta',
               'nolearn.lasagne.BatchIterator',
               'lasagne.layers',
               'keras.optimizers.SGD',
               'keras.utils.generic_utils',
               'lasagne.nonlinearities.softmax',
               'keras.regularizers.activity_l2',
               'lasagne.layers.Conv1DLayer',
               'keras.layers.core.AutoEncoder',
               'keras.layers.containers',
               'keras.callbacks.EarlyStopping',
               'keras.optimizers.Adagrad',
               'sklearn.neural_network.MLPClassifier',
               'lasagne.nonlinearities.rectify',
               'sklearn.neural_network.BernoulliRBM',
               'nolearn.lasagne.visualize',
               'lasagne.init.Uniform',
               'lasagne.updates.adagrad',
               'theano.tensor.nnet.conv'],
 'RandomForest': ['sklearn.ensemble.RandomForestClassifier',
                  'sklearn.ensemble.RandomForestRegressor'],
 'SVM': ['sklearn.svm.SVC',
         'sklearn.svm',
         'sklearn.svm.LinearSVC',
         'sklearn.svm.OneClassSVM']}
```

C

RÉSEAUX DE CO-OCCURENCES - *STACKEXCHANGE*

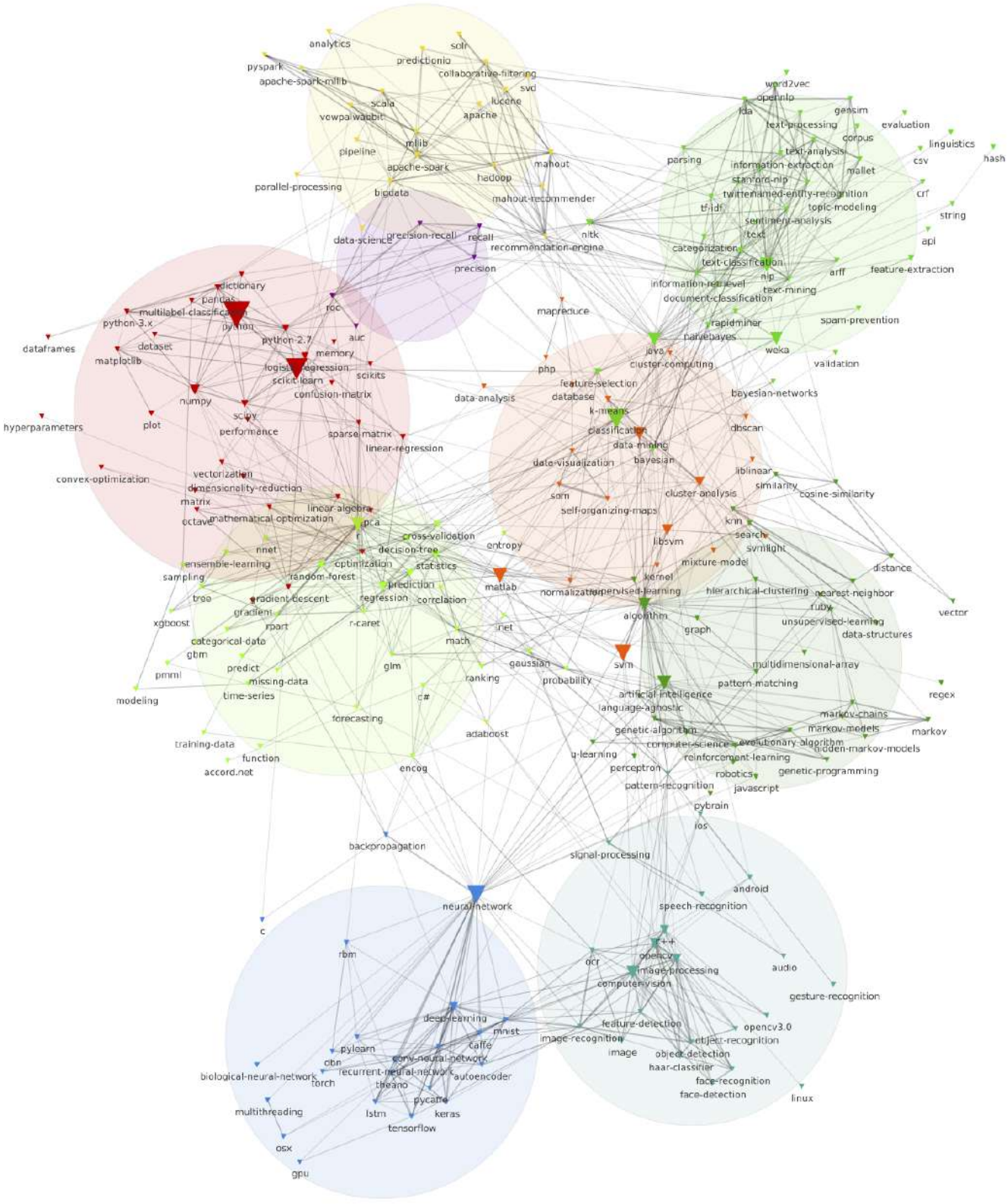


FIGURE 46 – Réseau de tags associés à l'apprentissage sur le site *stackoverflow*

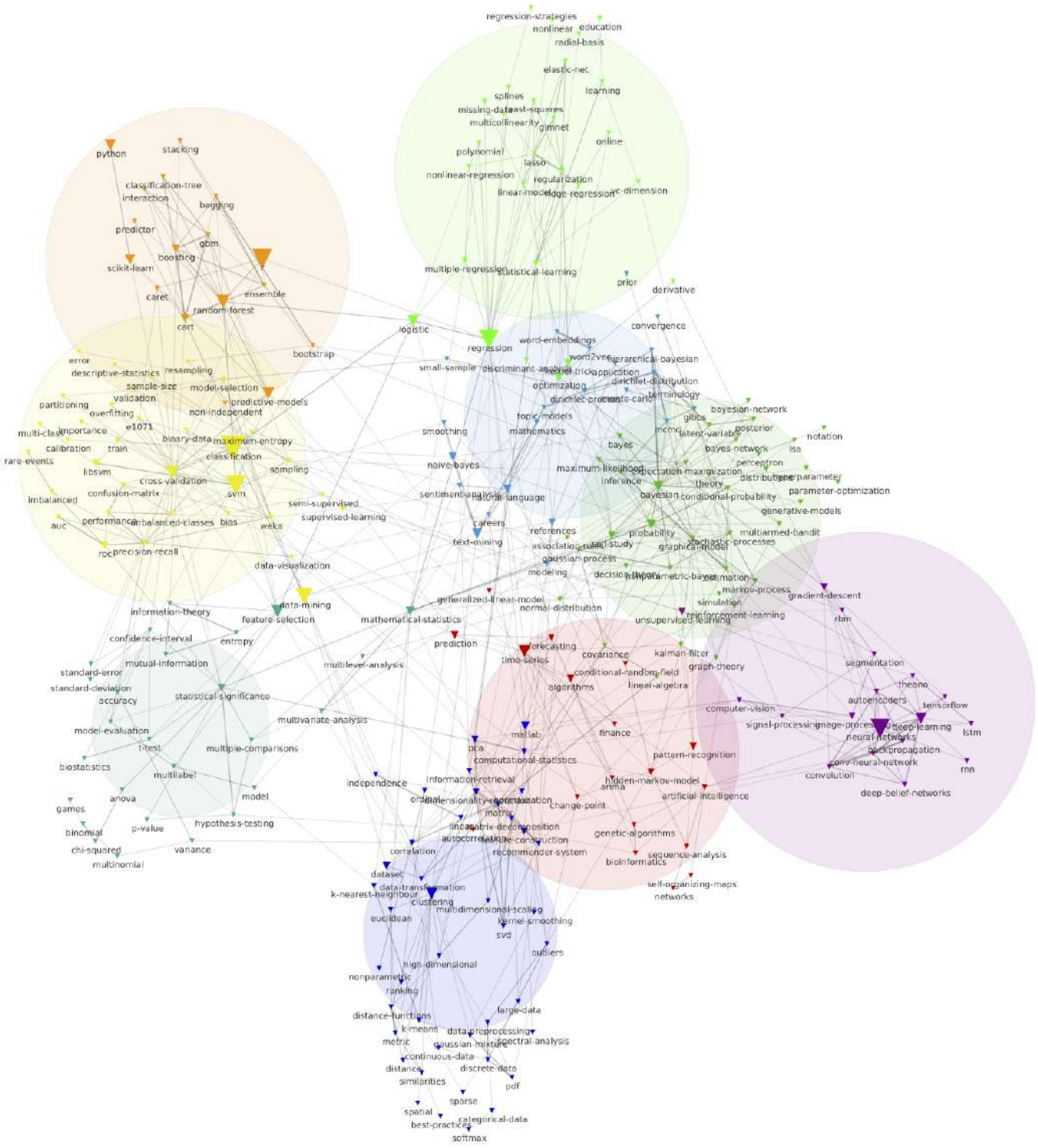


FIGURE 47 – Réseau de tags associés à l'apprentissage sur le site *Cross-Validated*

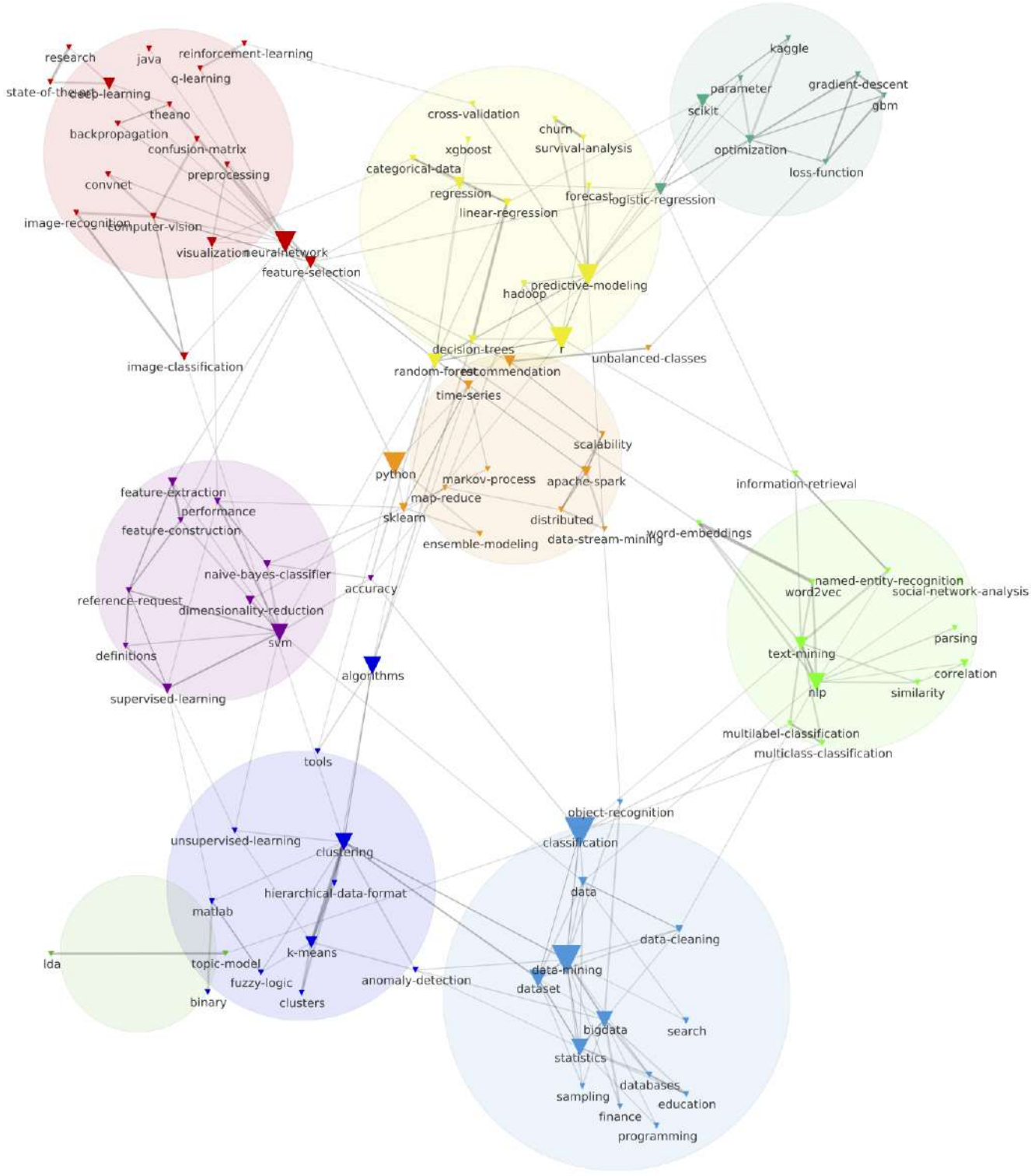


FIGURE 48 – Réseau de tags associés à l'apprentissage sur le site *stackoverflow*

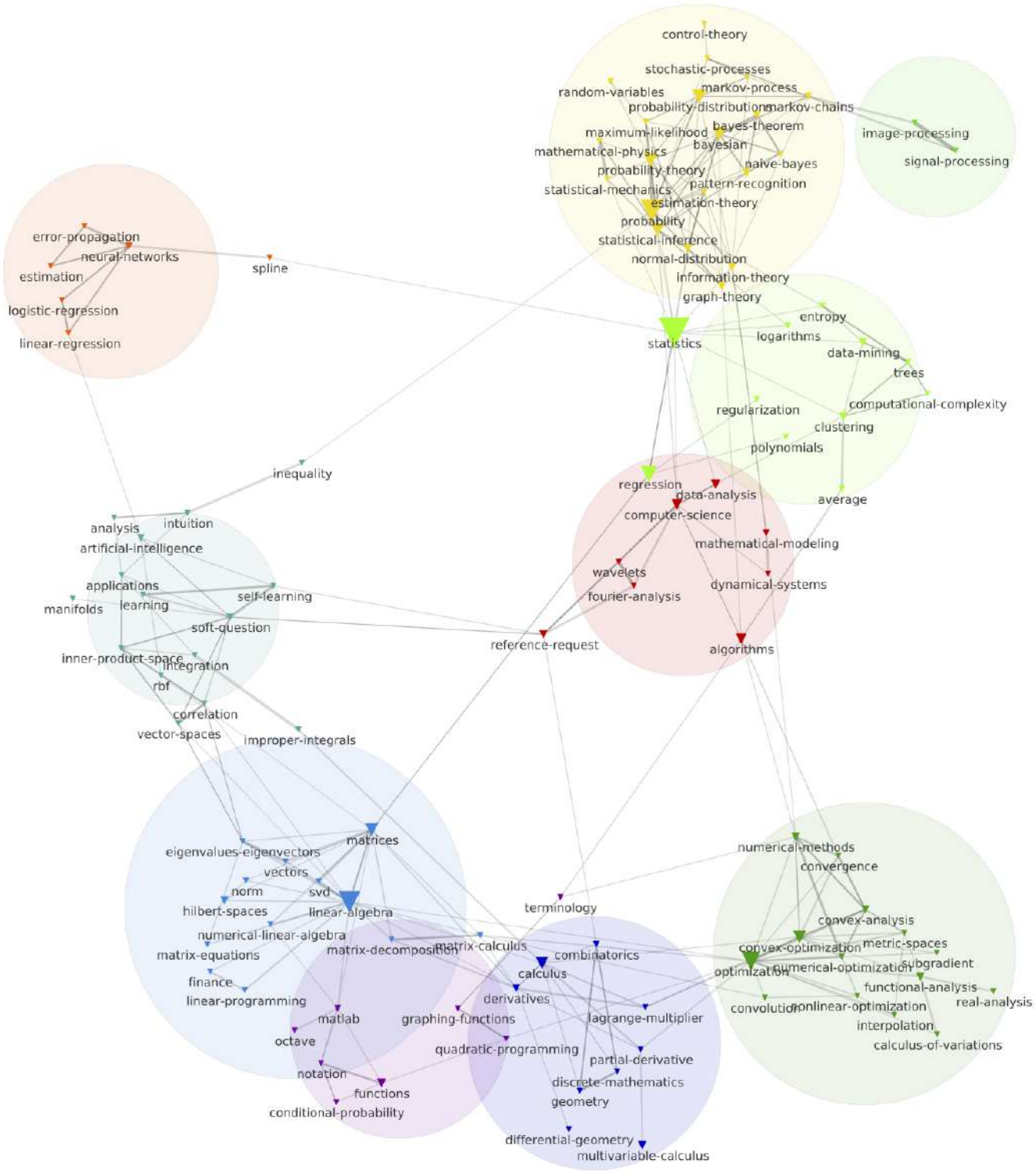


FIGURE 49 – Réseau de tags associés à l'apprentissage sur le site *stackoverflow*

BIBLIOGRAPHIE

[1] Scott Aaronson. *Quantum computing since Democritus*. Cambridge University Press, 2013.

[2] Scott Aaronson. *Is machine learning currently overhyped ?* 2015. URL : https://www.quora.com/Is-machine-learning-currently-overhyped/answer/Scott-Aaronson.

[3] Yaser S Abu-Mostafa, Malik Magdon-Ismail et Hsuan-Tien Lin. *Learning from data*. AMLBook Berlin, Germany, 2012.

[4] Yaser Abu-Mostafa. *Learning from data*. EdX. 2012.

[5] A Aizerman, Emmanuel M Braverman et LI Rozoner. « Theoretical foundations of the potential function method in pattern recognition learning ». In : *Automation and remote control* 25 (1964), p. 821–837.

[6] Chris Anderson. *The end of theory : The data deluge makes the scientific method obsolete*. 2008.

[7] John Austin. *Lectures on Jurisprudence : Or, The Philosophy of Positive Law*. 1875.

[8] Robert Axelrod. « Modeling the evolution of norms ». In : *American Political Science Association Meeting, New Orleans, LA*. 1985.

[9] Joseph Berkson. « Application of the logistic function to bio-assay ». In : *Journal of the American Statistical Association* 39.227 (1944), p. 357–365.

[10] Arthur S Bickel et Riva Wenig Bickel. « Tree structured rules genetic algorithms ». In : *Genetic algorithms and their applications : proceedings of the second International Conference on Genetic Algorithms : July 28-31, 1987 at the Massachusetts Institute of Technology, Cambridge, MA*. Hillsdale, NJ : L. Erlhaum Associates, 1987. 1987.

[11] Jakob Friedrich Freiherr von Bielfeld et William Hooper. *The Elements of Universal Erudition*. 1770.

[12] Muhammad Bilal Zafar. *Fair classification*. 2016. URL : https://github.com/mbilalzafar/fair-classification/.

[13] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez et Krishna P Gummadi. « Learning Fair Classifiers ». In : *arXiv preprint arXiv :1507.05259* (2015).

[14] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre. « Fast unfolding of communities in large networks ». In : *Journal of statistical mechanics : theory and experiment* 2008.10 (2008), P10008.

[15] Bernhard E Boser, Isabelle M Guyon et Vladimir N Vapnik. « A training algorithm for optimal margin classifiers ». In : *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, p. 144–152.

[16] Amiangshu Bosu, Christopher S Corley, Dustin Heaton, Debarshi Chatterji, Jeffrey C Carver et Nicholas A Kraft. « Building reputation in stackoverflow : an empirical investigation ». In : *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press. 2013, p. 89–92.

[17] Léon Bottou. « Two Big Challenges of Big Data ». In : *Keynote at International Conference of Machine Learning*. 2015.

[18] Leo Breiman. « Random forests ». In : *Machine learning* 45.1 (2001), p. 5–32.

[19] Leo Breiman, Jerome Friedman, Charles J Stone et Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[20] Leo Breiman et al. « Statistical modeling : The two cultures (with comments and a rejoinder by the author) ». In : *Statistical Science* 16.3 (2001), p. 199–231.

[21] Toon Calders et Sicco Verwer. « Three naive Bayes approaches for discrimination-free classification ». In : *Data Mining and Knowledge Discovery* 21.2 (2010), p. 277–292.

[22] Dominique Cardon. « Dans l'esprit du PageRank ». In : *Réseaux* 1 (2013), p. 63–95.

[23] Dominique Cardon. *A quoi rêvent les algorithmes : Nos vies à l'heure des big data*. Seuil, 2015.

[24] William S Cleveland. « Data science : an action plan for expanding the technical areas of the field of statistics ». In : *International statistical review* 69.1 (2001), p. 21–26.

[25] N Council. *Frontiers in massive data analysis*. 2013.

[26] David R Cox. « The regression analysis of binary sequences ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), p. 215–242.

[27] Charles Darwin. *L'origine des espèces*. 1859.

[28] Amit Datta, Michael Carl Tschantz et Anupam Datta. « Automated experiments on Ad privacy settings ». In : *Proceedings on Privacy Enhancing Technologies* 2015.1 (2015), p. 92–112.

[29] Gerald W Davis Jr. « Sensitivity analysis in neural net solutions ». In : *Systems, Man and Cybernetics, IEEE Transactions on* 19.5 (1989), p. 1078–1082.

[30] Richard Dawkins. *The selfish gene*. 199. Oxford university press, 2006.

[31] Stanislas Dehaene. *Le Code de la conscience*. Odile Jacob, 2014.

[32] Pedro Domingos. « A few useful things to know about machine learning ». In : *Communications of the ACM* 55.10 (2012), p. 78–87.

[33] Pedro Domingos. *The master algorithm*. Basic Civitas Books, 2015.

[34] Richard O Duda, Peter E Hart et al. *Pattern classification and scene analysis*. T. 3. Wiley New York, 1973.

[35] Ronald A Fisher. « The use of multiple measurements in taxonomic problems ». In : *Annals of eugenics* 7.2 (1936), p. 179–188.

[36] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[37] Nir Friedman, Dan Geiger et Moises Goldszmidt. « Bayesian network classifiers ». In : *Machine learning* 29.2-3 (1997), p. 131–163.

[38] Gerd Gigerenzer et Theodore Porter. *The empire of chance : How probability changed science and everyday life*. T. 12. Cambridge University Press, 1990.

[39] CW Gini. « Variability and mutability, contribution to the study of statistical distribution and relaitons ». In : *Studi Economico-Giuricici della R* (1912).

[40] Michelle Girvan et Mark EJ Newman. « Community structure in social and biological networks ». In : *Proceedings of the national academy of sciences* 99.12 (2002), p. 7821–7826.

[41] David E Goldberg et al. *Genetic algorithms in search optimization and machine learning*. T. 412. Addison-wesley Reading Menlo Park, 1989.

[42] Jerrold W Grossman. « The evolution of the mathematical research collaboration graph ». In : *Congressus Numerantium* (2002), p. 201–212.

[43] Benjamin V Hanrahan, Gregorio Convertino et Les Nelson. « Modeling problem difficulty and expertise in stackoverflow ». In : *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*. ACM. 2012, p. 91–94.

[44] Yuval Harari. *Techno-Religions and Silicon Prophets*. 2015. url : https://www.youtube.com/watch?v=g6BK5Q_Dblo.

[45] Yuval Harari. *Homo Deus : A brief history of tomorrow*. Harvill Secker, 2016.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. « Deep Residual Learning for Image Recognition ». In : *arXiv preprint arXiv :1512.03385* (2015).

[47] Geoffrey Hinton. *Neural Nets for Machine Learning*. Coursera. 2012.

[48] Douglas R Hofstadter. « Godel escher bach ». In : *New Society* (1980).

[49] John H Holland. *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[50] White House. *Big data : Seizing opportunities, preserving values*. 2014.

[51] White House. *Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights*. 2016.

[52] Orna Intrator et Nathan Intrator. « Interpreting neural-network results : a simulation study ». In : *Computational statistics & data analysis* 37.3 (2001), p. 373–393.

[53] Anil K Jain, M Narasimha Murty et Patrick J Flynn. « Data clustering : a review ». In : *ACM computing surveys (CSUR)* 31.3 (1999), p. 264–323.

[54] Faisal Kamiran, Toon Calders et Mykola Pechenizkiy. « Discrimination aware decision tree learning ». In : *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE. 2010, p. 869–874.

[55] Andrej Karpathy. *The unreasonable effectiveness of recurrent neural networks*. 2015. URL : http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

[56] Katty Kay et Claire Shipman. « When Discrimination Is Baked Into Algorithms ». In : *The Atlantic* (2014).

[57] Kevin Kelly. *The Three Breakthroughs That Have Finally Unleashed AI on the World*. 2014. URL : http://www.wired.com/2014/10/future-of-artificial-intelligence/.

[58] Madian Khabsa et C Lee Giles. « The number of scholarly documents on the public web ». In : *PloS one* 9.5 (2014), e93949.

[59] Lauren Kirchner. « When Discrimination Is Baked Into Algorithms ». In : *The Atlantic* (2015).

[60] Scott Kirkpatrick. « Optimization by simulated annealing : Quantitative studies ». In : *Journal of statistical physics* 34.5-6 (1984), p. 975–986.

[61] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems*. 2012, p. 1097–1105.

[62] Pierre-Simon Laplace. *Mémoire sur la Probabilité des Causes par les Événements*. 1774.

[63] Pierre-Simon Laplace. *Exposition du système du monde*. 1796.

[64] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann et al. « Life in the network : the coming age of computational social science ». In : *Science* 323.5915 (2009), p. 721.

[65] Quoc V Le. « Building high-level features using large scale unsupervised learning ». In : *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, p. 8595–8598.

[66] Yann LeCun. « Une procedure d'apprentissage pour reseau a seuil asymmetrique (A learning scheme for asymmetric threshold networks) ». In : (1985).

[67] Yann LeCun, Yoshua Bengio et Geoffrey Hinton. « Deep learning ». In : *Nature* 521.7553 (2015), p. 436–444.

[68] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard et Lawrence D Jackel. « Backpropagation applied to handwritten zip code recognition ». In : *Neural computation* 1.4 (1989), p. 541–551.

[69] Yann LeCun, Léon Bottou, Yoshua Bengio et Patrick Haffner. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11 (1998), p. 2278–2324.

[70] Yann Lecun. « What's wrong with deep learning ? » In : *CPVR15*. 2015.

[71] Curt A Levey. « Neural network having expert system functionality ». Brev. 5,398,300. 1995.

[72] John Markoff. « What's the best answer ? It's survival of the fittest ». In : *New York Times* (1990).

[73] Trent McConaghy. « Ffx : Fast, scalable, deterministic symbolic regression technology ». In : *Genetic Programming Theory and Practice IX*. Springer, 2011, p. 235–260.

[74] Pamela McCorduck. « Machines who think ». In : (1979).

[75] Warren S McCulloch et Walter Pitts. « A logical calculus of the ideas immanent in nervous activity ». In : *The bulletin of mathematical biophysics* 5.4 (1943), p. 115–133.

[76] Sharon Bertsch McGrayne. *The theory that would not die*. Yale University Press, 2011.

[77] Claire Cain Miller. « When algorithms discriminate ». In : *New York Times* 9 (2015).

[78] John Mingers. « Rule induction with statistical data-a comparison with multiple regression ». In : *Journal of the operational research Society* (1987), p. 347–351.

[79] Marvin Minsky et Seymour Papert. « Perceptrons. » In : (1969).

[80] Tom Mitchell. *Machine Learning*. McGraw Hill, 1998.

[81] Abraham de Moivre. *The Doctrine of Chances*. 1718.

[82] Edward C Molina. « Translating and selecting system. » Brev. 1,083,456. 1914.

[83] Seyed Mehdi Nasehi, Jonathan Sillito, Frank Maurer et Chris Burns. « What makes a good code example ? : A study of programming Q&A in StackOverflow ». In : *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*. IEEE. 2012, p. 25–34.

[84] Peter Naur. « The place of programming in a world of problems, tools, and people ». In : *Proceedings of the IFIP Congress*. T. 65. 1965, p. 195–199.

[85] Peter Naur. *Concise survey of computer methods*. Petrocelli Books, 1974.

[86] Allen Newell. *Intellectual issues in the history of artificial intelligence*. Rapp. tech. 1982.

[87] Mark EJ Newman. « Coauthorship networks and patterns of scientific collaboration ». In : *Proceedings of the national academy of sciences* 101.suppl 1 (2004), p. 5200–5205.

[88] Mark EJ Newman. « Fast algorithm for detecting community structure in networks ». In : *Physical review E* 69.6 (2004), p. 066133.

[89] Isaac Newton. *Principes mathématiques de la philosophie naturelle*. 1687.

[90] Andrew Ng. *Machine Learning*. Coursera. 2011.

[91] Mikel Olazaran. « A sociological study of the official history of the perceptrons controversy ». In : *Social Studies of Science* 26.3 (1996), p. 611–659.

[92] Peggy Orenstein. *Schoolgirls : Young women, self esteem, and the confidence gap*. Anchor, 2013.

[93] Lawrence Page, Sergey Brin, Rajeev Motwani et Terry Winograd. « The PageRank citation ranking : bringing order to the web. » In : (1999).

[94] Judea Pearl. *Causality*. Cambridge university press, 2009.

[95] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, 2014.

[96] Karl Pearson. « La grammaire de la science ». In : *Journal de la société de statistique de Paris.* 53 (1912), p. 196–214.

[97] Karl Pearson. « Laplace ». In : *Biometrika* (1929).

[98] François Petitjean, Geoffrey I Webb et Ann E Nicholson. « Scaling log-linear analysis to high-dimensional data ». In : *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE. 2013, p. 597–606.

[99] Henry C Plotkin. *Darwin machines and the nature of knowledge*. Harvard University Press, 1997.

[100] Karl Popper. « Objective knowledge : An evolutionary approach ». In : (1972).

[101] J. Ross Quinlan. « Induction of decision trees ». In : *Machine learning* 1.1 (1986), p. 81–106.

[102] J Ross Quinlan. *C4.5 : programs for machine learning*. Elsevier, 2014.

[103] Gerard Radnitzky, William Warren Bartley et Karl Raimund Popper. *Evolutionary epistemology, rationality, and the sociology of knowledge*. Open Court Publishing, 1987.

[104] Benjamin Raimbault, Jean-Philippe Cointet et Pierre-Benoit Joly. « Analyse de l'émergence de la biologie de synthèse : Une approche scientométrique ». In : *Troisième colloque Sciences de la vie en société*. Génolopole, ifris. 2014.

[105] Gunnar Rätsch, Sören Sonnenburg et Christin Schäfer. « Learning interpretable SVMs for biological sequence classification ». In : *BMC bioinformatics* 7.Suppl 1 (2006), S9.

[106] George Rebane et Judea Pearl. « The recovery of causal poly-trees from statistical data ». In : *arXiv preprint arXiv :1304.2736* (2013).

[107] Ingo Rechenberg. *Evolutionsstrategie*. Stuttgart : Holzmann-Froboog, 1973.

[108] R Riolo. « CFS-C : A Package of Domain Independent Subroutines for Implementing Classifier Systems in Arbitrary User-Defined Environments ». In : *Logic of Computers Group, Division of Computer Science and Engineering, University of Michigan* (1986).

[109] Anna W Roe, Sarah L Pallas, Young H Kwon et Mriganka Sur. « Visual projections routed to the auditory pathway in ferrets : receptive fields of visual neurons in primary auditory cortex ». In : *The Journal of neuroscience* 12.9 (1992), p. 3651–3664.

[110] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[111] Frank Rosenblatt. « The perceptron : a probabilistic model for information storage and organization in the brain. » In : *Psychological review* 65.6 (1958), p. 386.

[112] Frank Rosenblatt. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Rapp. tech. DTIC Document, 1961.

[113] Alix Rule, Jean-Philippe Cointet et Peter S Bearman. « Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014 ». In : *Proceedings of the National Academy of Sciences* 112.35 (2015), p. 10837–10844.

[114] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. « Learning representations by back-propagating errors ». In : *Cognitive modeling* 5.3 (1988), p. 1.

[115] Stuart Russell et Peter Norvig. *A modern approach*. Pearson, 1995.

[116] Oliver Sacks. *L'homme qui prenait sa femme pour un chapeau*. Seuil, 1988.

[117] Arthur L Samuel. « Some studies in machine learning using the game of checkers ». In : *IBM Journal of research and development* 3.3 (1959), p. 210–229.

[118] Michael Schmidt et Hod Lipson. « Distilling free-form natural laws from experimental data ». In : *science* 324.5923 (2009), p. 81–85.

[119] Philip A Schrodt. « Predicting international events ». In : *Byte* 11.12 (1986), p. 177–192.

[120] Hans-Paul Schwefel. *Numerische optimierung von computer-modellen mittels der evolutionsstrategie*. T. 1. Birkhäuser, Basel Switzerland, 1977.

[121] Galit Shmueli. « To explain or to predict ? » In : *Statistical science* (2010), p. 289–310.

[122] D Simberloff, BC Barish, KK Droegemeier, DM Etter, NV Fedoroff, KM Ford, LJ Lanzerotti, A Leshner, J Lubchenco, MG Rossmann et al. *Long-lived digital data collections : enabling research and education in the 21st century*. Rapp. tech. Technical Report NSB-05-40, NSF, USA, 2005.

[123] Henry Small. « Co-citation in the scientific literature : A new measure of the relationship between two documents ». In : *Journal of the American Society for information Science* 24.4 (1973), p. 265–269.

[124] Stephen Frederick Smith. « A learning system based on genetic adaptive algorithms ». In : (1980).

[125] Joel Spolsky. *The Cultural Anthropology of Stack Exchange*. https://www.youtube.com/watch?v=LpGA2fmAHvM. 2012.

[126] Clayton Stanley et Michael D Byrne. « Predicting tags for stackoverflow posts ». In : *Proceedings of ICCM*. T. 2013. 2013.

[127] Stephen M Stigler. « Who discovered Bayes's theorem ? » In : *The American Statistician* 37.4a (1983), p. 290–296.

[128] Stephen M Stigler. *The history of statistics : The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[129] William Stukeley. *Memoirs of Sir Isaac Newton's Life*. 1752.

[130] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich. « Going deeper with convolutions ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, p. 1–9.

[131] Jason Tanz. *The rise of Artificial Intelligence and the end of code*. 2016.

[132] Joshua B Tenenbaum, Thomas L Griffiths et Charles Kemp. « Theory-based Bayesian models of inductive learning and reasoning ». In : *Trends in cognitive sciences* 10.7 (2006), p. 309–318.

[133] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths et Noah D Goodman. « How to grow a mind : Statistics, structure, and abstraction ». In : *science* 331.6022 (2011), p. 1279–1285.

[134] B Thompson. « Evolving knowledge from data ». In : *Computer Language* 3.11 (1986), p. 23–26.

[135] John W Tukey. « The future of data analysis ». In : *The Annals of Mathematical Statistics* 33.1 (1962), p. 1–67.

[136] John W Tukey. « Exploratory data analysis ». In : (1977).

[137] Alan M Turing. « Computing machinery and intelligence ». In : *Mind* 59.236 (1950), p. 433–460.

[138] Leslie Valiant. *Probably Approximately Correct : Natures Algorithms for Learning and Prospering in a Complex World*. Basic Books, 2013.

[139] V Vapnik et A Lerner. « Pattern recognition using generalized portrait method ». In : *Automation and remote control* 24 (1963), p. 774–780.

[140] Vladimir Vapnik. *The nature of statistical learning theory*. T. 1. Springer, 1995.

[141] Vladimir Vapnik et A Chervonenkis. « On the uniform convergence of relative frequencies of events to their probabilities ». In : *Measures of Complexity*. Springer, 2015, p. 11–30.

[142] Bogdan Vasilescu, Andrea Capiluppi et Alexander Serebrenik. « Gender, representation and online participation : A quantitative study of stackoverflow ». In : *Social Informatics (SocialInformatics), 2012 International Conference on*. IEEE. 2012, p. 332–338.

[143] Voltaire. *Éléments de la Philosophie de Newton*. 1738.

[144] Shaowei Wang, David Lo et Lingxiao Jiang. « An empirical study on developer interactions in StackOverflow ». In : *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM. 2013, p. 1019–1024.

[145] Paul Werbos. « Beyond regression : New tools for prediction and analysis in the behavioral sciences ». In : (1974).

[146] Mark J Willis, Hugo G Hiden, B McKay, GA Montague et P Marenbach. « Genetic programming : An introduction and survey of applications ». In : *IEE conference publication*. Institution of Electrical Engineers. 1997, p. 314–319.

[147] Jeff Wu. « Statistics = Data Science ? » In : *Inaugural lecture for the Carver Chair at the University of Michigan*. 1997.

[148] Fei Xu et Joshua B Tenenbaum. « Word learning as Bayesian inference. » In : *Psychological review* 114.2 (2007), p. 245.

[149] Wojciech Zaremba et Ilya Sutskever. « Learning to execute ». In : *arXiv preprint arXiv :1410.4615* (2014).

[150] Harry Zhang. « The optimality of naive Bayes ». In : *AA* 1.2 (2004), p. 3.

[151] Yangyong Zhu et Yun Xiong. « Towards Data Science ». In : *Data Science Journal* 14 (2015).