Optimisation et Science des Données : entre théorie et applications

Stefan Janaqi

Mémoire

En vue de la constitution d'un dossier de candidature à l'HDR

Stefan Janaqi
IMT Mines-Alès
Euromov DHM
Mai 2019

Table des matières

0. Curriculum Vitae

En vue de la constitution d'un dossier de candidature à l'Habilitation à Diriger des Recherches
Université de Montpellier – Ecole doctorale I2S – Spécialité Informatique

---

Stefan Janaqi
Né le 25 avril 1960 à Korçe (Albanie) – Nationalité française – Marié – 2 enfants

Maître Assistant à l'IMT – Mines Alès
Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P)
IMT Mines Alès - Bâtiment M-7, rue Jules Renard - 30319 Alès
Email : stefan.janaqi@mines-ales.fr – Téléphone : +33 (0)6 52 97 27 73

Parcours Professionnel

Depuis janvier 2020
Membre de l'UMR EuroMov Digital Health in Motion – Université de Montpellier, IMT Mines Alès.

Depuis janvier 2018
Membre de l'équipe dédiée TIC & Santé du LGI2P accueillie au centre de recherche EuroMov de l'Université de Montpellier.

Depuis octobre 2000 Enseignant-Chercheur
IMT – Mines Alès – titulaire depuis octobre 2001.

Octobre 1997 – octobre 2000. Ingénieur en Mathématiques Appliquées
Société ELF – Antar France. Société Décan.

Octobre 1995 – octobre 1997. Post Doctorat en Mathématiques Appliquées
Société ELF Antar France.

Diplômes

1995 Doctorat (Ph. D.) – Université Joseph Fourier, Grenoble
Laboratoire des Structures Discrètes et Didactique (IMAG)
Spécialité : Informatique, option Recherche Opérationnelle
Titre de la thèse : Quelques éléments de la géométrie des graphes
Directeur de thèse : Charles Payan (CNRS)
Soutenue le : 27-9-1995 à Grenoble – Mention : Très Honorable avec les Félicitations du jury.

1991 D.E.A. (M. Sc.) – Université Joseph Fourier, Grenoble

1. Activités de recherche et d'encadrement de chercheurs

Je suis membre de l'UMR EuroMov Digital Health in Motion – Université de Montpellier et IMT Mines Alès (EuroMov DHM). Cette unité a été créée en janvier 2020. Elle a résulté de la fusion du laboratoire EuroMov de l'Université de Montpellier et d'une grande partie de chercheurs du LGI2P – IMT Mines Alès, notamment ceux de l'équipe KID « Knowledge representation and Image analysis for Decision » dont je faisais partie. Un des objectifs de cette équipe est le développement de méthodes de représentation, de transformation et d'analyse de données afin d'extraire l'information pertinente par des méthodes d'apprentissage automatique. Ces procédures de recherche d'information s'appuient sur une large palette d'outils mathématiques, méthodes d'optimisation, théorie des graphes, théorie de l'information, statistiques, logiques de représentation des connaissances, techniques de traitement du signal, méthodes numériques. Mon parcours d'études universitaires, doctorales et postdoctorales m'a permis d'avoir un ensemble de connaissances qui trouvent une place légitime dans la palette mentionnée. Un autre élément aussi important de mon parcours est une expérience industrielle de cinq ans qui a forgé le lien entre le savoir académique, les concepts, les modèles mathématiques d'un côté et la mise en pratique dans des milieux réels avec des contraintes qui dépassent de loin les cadres théoriques des modèles. Cette activité croisée et productive entre les concepts fondamentaux théoriques et les applications du monde réel, activité qui m'a suivi durant mon parcours professionnel, est la caractéristique principale de mon profil.

Dans le cadre d'EuroMov DHM, je suis chargé de suivre des études et des projets faisant appel aux mathématiques appliquées. Pratiquement, une grande partie de mes activités de recherche, d'encadrements de sujets de thèse, d'enseignements ainsi que les projets industriels, peut se projeter sur les deux axes principaux suivants :

(i) Optimisation ;
(ii) Apprentissage artificiel ;

Ces deux axes ne sont pas indépendants, ils communiquent et s'enrichissent mutuellement et puisent dans un socle commun de concepts mathématiques et informatiques sur lesquels je me suis formé tout au long de mon cursus d'études et surtout de mon activité professionnelle.

Optimisation

Il m'arrive souvent pendant mes enseignements de la théorie des graphes d'expliquer aux étudiants que la plupart des « bons » algorithmes cherchent une information globale à partir d'informations locales. Ce fait remarquable est central pour de grandes classes de problèmes et applications. L'exemple indiqué est l'algorithme du plus court chemin.

Cette possibilité d'utilisation si « efficace » de l'information locale repose sur des structures fondamentales présentes dans les objets mathématiques continus ou discrets. Parmi ces structures la

convexité est une notion géométrique importante en mathématiques et largement utilisée en optimisation continue. La raison principale du succès de la convexité tient au fait que pour un problème d'optimisation convexe un extrémum local est global.

Cette notion de convexité n'est pas nécessairement liée aux espaces à support continu et peut aussi bien se définir dans un graphe ou dans un ensemble discret. L'utilisation de la combinatoire pour étudier la convexité dans les espaces euclidiens a connu des succès remarquables dans l'étude de la structure des polytopes et de leur impact en optimisation. Les algorithmes de simplexe, point intérieur, d'optimisation convexe en sont des illustrations remarquables.

Qu'en est-il de l'utilisation de la notion de convexité en combinatoire ? Cette facette combinatoire de la convexité est moins étudiée. Les travaux de mon doctorat intitulé « Quelques éléments de la géométrie des graphes » (voir publication 43) se sont focalisés dans cette direction. Il s'agit de connaitre les mécanismes mathématiques fondamentaux qui définissent la convexité. Il est possible de définir axiomatiquement une structure convexe à partir d'un opérateur de fermeture. De façon plus concrète, un plus court chemin, ou plutôt l'ensemble de plus courts chemins reliant deux sommets d'un graphe joue le rôle analogue d'un segment de droite dans l'espace euclidien. Cet ensemble de plus courts chemins est appelé intervalle. L'action répétée de cet opérateur intervalle sur un ensemble de sommets de départ finit par s'arrêter fournissant ainsi un ensemble convexe. A partir de là, plusieurs concepts reliés à la convexité peuvent être définis et employés pour caractériser des propriétés structurelles des graphes. La notion d'ensemble générateur est fondamentale. Un ensemble minimal de sommets peut générer par une opération de fermeture convexe une partie ou le graphe entier (voir liste de publications 12, 13, 16, 43). La fermeture convexe est une façon de créer un ensemble convexe. La deuxième façon est la création d'un ensemble convexe par coupes successives. Ces deux approches sont au cœur d'un des concepts fondamentaux : la dualité. C'est un des concepts les plus importants tant d'un point de vue théorique que pratique, car il fournit des preuves théoriques et des algorithmes d'optimisation efficaces.

La dualité a son mot à dire dans le cas discret aussi. Une contribution majeure de ma thèse est la preuve d'un résultat min-max sur une classe de graphes généralisant les polyminoes : la cardinalité minimum d'un générateur convexe est égale à la cardinalité maximum des sommets pendants d'un arbre obtenu à partir du graphe initial par une suite de contractions successives de coupes de Djokovitch.

En plus de son intérêt théorique, cette approche axiomatique de la convexité a rencontré une application directe quelques années plus tard. Lorsqu'un ensemble de concepts reliés par une relation (e.g. lion $is-a$ mammifère) est organisé en ontologie, il est nécessaire de calculer le plus petit ensemble contenant une liste initiale de concepts de base. Ce plus petit ensemble est défini par rapport à la fermeture de l'opérateur $\varphi \equiv is-a$ qui vérifie les quatre axiomes d'une structure convexe $\mathbb{C}$ définie sur un ensemble E:

(a1) $\emptyset \in \mathbb{C}, E \in \mathbb{C}, \varphi(\emptyset) = \emptyset, \varphi(E) = E$;
(a2) (monotonie) $A \subseteq B \rightarrow \varphi(A) \subseteq \varphi(B)$;
(a3) (idempotence) $\varphi\big(\varphi(A)\big) = \varphi(A)$;
(a4) (générateur fini) $x \in A \rightarrow \exists\, F \subseteq A$ tel que $x \in \varphi(F)$ et $card(F) < \infty$.

Ce petit ensemble d'axiomes contient la clé pour concevoir un algorithme efficace de génération d'ensembles de concepts complets pour l'opérateur $is-a$ (voir publications 9 et 10).

La convexité implique aussi des propriétés structurelles d'un graphe. En interdisant les deux graphes suivants d'être des sous graphes convexes d'un graphe G, on peut prouver que G est un produit cartésien d'arbres et/ou chemins. Ce résultat (voir publication 43) a été la preuve d'une conjecture de Burosch et Laborde (Burosch G. Laborde J.-M., Characterization of grid graphs, Discrete Mathematics, 87, 1991, p 85-88).

Figure 1. Deux sous graphes interdits minimaux non convexes. Un graphe est un produit de chemins ou d'arbres s'il n'a pas de sous-graphe convexe isomorphe à un des deux graphes ci-dessus.

Ces travaux illustrent le versant théorique de mon profil à la recherche de structures et de conditions mathématiques minimales (locales) qui fournissent des propriétés structurelles (globales). Je me suis servi à plusieurs reprises de ce versant pour concevoir des approches et des solutions autant dans le contexte de travaux de recherche et l'enseignement que dans la résolution de problèmes d' applications industrielles.

À la suite de mon doctorat, j'ai eu mon premier contrat post-doctoral dans un centre de recherche de la société ELF Antar France. Le sujet était : A partir d'un stock de bases chimiques donné, maximiser la probabilité de produire le plus de mélanges possibles. Ce fut une plongée passionnante au cœur de problèmes d'optimisation dans un cadre industriel

J'ai acquis les bases mathématiques de l'optimisation et de la recherche opérationnelle durant ma formation en mathématiques appliquées à la Faculté des Mathématiques de l'Université de Tirana en Albanie. Par suite de mes études, j'ai enseigné l'optimisation et la théorie des graphes en tant que Maître Assistant à la Faculté des Sciences de l'Université de Tirana.

Tout un monde sépare la théorie de l'optimisation dont les problèmes vérifient toutes les propriétés souhaitées de la pratique de l'optimisation où les données sont incertaines, la précision des processeurs est finie et les cas dégénérés arrivent souvent.

Une modélisation mathématique intéressante du problème posé utilise les zonotopes (c'est un corps convexe particulier qui peut être décrit comme somme de Minkowski de segments de droite ou image affine d'un hypercube), décrivant l'ensemble de mélanges réalisables. Le volume du zonotope peut être calculé explicitement et la racine $n^{ième}$ (n étant la dimension de l'espace) de ce volume est une fonction concave et donc avec un unique maximum. La solution du problème posé était de fabriquer le mélange courant (tous les mélanges ne sont pas connus au préalable) de telle sorte que le volume du zonotope résiduel fut maximal. Le logiciel correspondant et le choix des solveurs ont répondu aux attentes industrielles. C'était une expérience fondamentale qui m'a définitivement tourné vers les applications des mathématiques.

Il y a donc un fossé important entre le cadre théorique de l'optimisation, où les fonctions vérifient une suite de propriétés qui assurent la preuve de résultats d'existence et de convergence, et la pratique de l'optimisation qui se heurte souvent à des problèmes difficiles. Souvent les fonctions à optimiser ne sont pas convexes, pas différentiables et même pas continues. Au mieux, on possède une routine

qui retourne la valeur de $f(x)$ pour un x donné. Une de mes contributions en optimisation est l'encadrement majeur d'une thèse sur l'évolution différentielle (thèse T1). Il s'agit de faire évoluer vers un optimum local un nuage de points candidats au lieu d'un seul point comme c'est le cas de plusieurs algorithmes classiques d'optimisation. Ce nuage peut être choisi initialement pour couvrir la zone d'intérêt. Les stratégies d'évolution de ce nuage sont basées sur le gain en valeurs de $f(x)$ d'où le nom de la méthode. Il est aisé de contraindre l'évolution des points selon des directions orthogonales afin d'éviter les cas dégénérés. L'évolution des points est réalisée par mouvement sur des lignes de droite ce qui permet de conserver facilement la faisabilité des solutions. Pour contrôler la qualité des solutions et fournir une condition d'arrêt, on se base sur une approximation d'ordre deux de la surface de réponse calculée à partir des couples $(x, f(x))$ du nuage. C'est une surface quadratique approximant les points par une régression support vector machines (SVM). Ainsi, parallèlement avec l'évolution différentielle du nuage, on apprend la surface de réponse et on suit son évolution au fur et à mesure des itérations. Le choix des SVM comme méthode d'apprentissage n'est pas anodin. Les SVM fournissent les support vectors, les points les plus proches de la surface ou de façon duale, les points dont les variables duales valent zéro, qui sont cruciaux pour la définition de la surface de réponse. Ensuite, ce sont ces vecteurs qui évolueront en priorité. La condition d'arrêt de l'évolution est basée sur la dégénérescence du Hessien qui correspond à la platitude locale de la surface de réponse. Un avantage important de l'évolution différentielle réside dans le fait que c'est une méthode nativement distribuée et donc facilement parallélisable (voir publications 33, 34, 35). Cette thèse allie les approches théoriques et conceptuelles avec les solutions pratiques efficaces sur des cas réels. Cette méthode a été testée avec succès sur des benchmarks de problèmes d'optimisation ainsi que sur des problèmes industriels.

Un autre aspect important en optimisation est la robustesse des résultats en présence de données incertaines, ce qui est le cas sur la majorité des applications industrielles. Les données qui déterminent les coefficients des contraintes et de la fonction objectif sont entachées d'erreurs. Il y a plusieurs sources d'erreur : précision limitée des appareils d'analyse ; erreurs humaines ; plusieurs facteurs exogènes ne sont pas pris en compte lors de la formulation du problème d'optimisation ; les données du problème évoluent avec le temps. Ces facteurs parmi d'autres font que la faisabilité et l'existence d'un optimum peuvent être compromises ou erronées même dans le cas « simple » de problèmes linéaires.

Typiquement, le problème linéaire nominal que l'on soumet au solveur n'est qu'un problème parmi un ensemble $\mathbf{P}$ de problèmes :

(i) (Problème Nominal) $minimiser\ f(x), Ax \leq b, A \in \mathbf{P}$;
(ii) (Formulation Robuste) $minimiser\ f(x), Ax \leq b,\ \forall A \in \mathbf{P}$;

La question se pose de savoir quel lien existe entre ces deux formulations. Si le problème (ii) est infaisable, existe-t-il une instance infaisable (i) ? Si (ii) est faisable et de valeur optimale f_{opt} existe-t-il une instance faisable (i) qui atteigne cette valeur optimale ? La réponse est « Non » pour ces deux questions comme prouvé par Ben-Tal et Nemirovskii (Robust solutions of uncertain linear programs, Ben-Tal A., Nemirovskii A., 1999-2003). Pour avoir deux réponses « Oui » les conditions suivantes sont suffisantes :

> Condition 1. L'ensemble $\mathbf{P_k}$ de variation de la ligne A_k de A est convexe pour toutes les contraintes $k = 1, \dots, m$ et $\mathbf{P} = \mathbf{P_1} \times \dots \times \mathbf{P_m}$ est leur produit cartésien.

Condition 2. Il existe un compact $\mathbf{K}$ contenant les zones faisables de tous les problèmes (ii).

Les données du problème permettent souvent de satisfaire la Condition 2 en fournissant des bornes min-max sur la variable d'optimisation $l \leq x \leq u$. La satisfaction de la Condition 1 nécessite de définir une zone de variation de chaque ligne A_k de A. Un cadre suffisamment général est de considérer l'ensemble de variation $A_k(z) = A_k^0 + z^T Q_k, \|z\|_p \leq 1, p \geq 1$. Autrement dit, $A_k(z)$ est l'image d'une boule fermée de rayon 1 pour la norme L_p. Classiquement la norme quadratique L_2 est utilisée. On peut montrer qu'une solution x est robuste (problème (ii)) si et seulement si $A_k^0 x + \|Q_k x\|_q \leq b_k, k = 1, \dots, m$. Ainsi, une solution robuste est strictement à l'intérieur de la zone faisable et ce degré « d'intériorité » est mesuré par la norme duale $L_q, \frac{1}{p}+\frac{1}{q} = 1$, $\|Q_k x\|_q$. Ceci est intuitivement clair, les solutions les plus à l'intérieur sont moins impactées par l'incertitude des frontières de la zone. Cette approche fut la base d'un encadrement majeur d'une thèse sur l'optimisation robuste en temps réel (thèse T2). Cette thèse était en lien avec des problématiques industrielles de mélanges. La contrainte de résoudre le problème, ou plutôt une suite de problèmes en temps réel fut un défi autant en termes de formulations mathématiques que de solutions logicielles. En s'écartant du cas classique de la norme quadratique L_2, un couple candidat de normes duales est L_1 et L_∞. Ces deux normes permettent de formuler le problème robuste comme un problème linéaire. Plusieurs décennies de développement des solveurs de problèmes linéaires ont permis de proposer des solutions logicielles robustes satisfaisant la contrainte temps réel. Un autre apport de cette thèse fut une aide à la décision par la visualisation de coupes 2D significatives pour indiquer aux opérateurs la position des solutions dans le polytope faisable. En plus des publications (voir publications 20, 21), les programmes issus de ces travaux tournent actuellement en H24.

L'optimisation de *loss functions* est une des pierres fondamentales des méthodes d'apprentissage automatique. La régression classique n'est que la minimisation d'une norme quadratique. Actuellement la masse de données cumulées augmente mais leur dimension aussi. Dans plusieurs situations la dimension (et sa « malédiction ») augmente bien plus vite que le nombre d'observations. Typiquement, dans les applications médicales, le nombre d'observations est limité par le nombre de personnes passant un test alors que le nombre de mesures recueillies peut facilement atteindre des ordres $10^4 - 10^6$. En même temps dans plusieurs cas il y a une grande redondance entre les variables. Typiquement, lorsque l'on marche la coordination du mouvement implique l'existence de la même information sur plusieurs points du corps humain. Dans ce cadre il s'agit de chercher un sous ensemble minimal de variables explicatives pour prédire la réponse. C'est un problème fondamentalement combinatoire. Des heuristiques de régularisation / pénalisation, basées sur la notion de dualité, fournissent des résultats excellents en éliminant un grand nombre de variables redondantes. Actuellement, je contribue à l'élaboration de techniques d'optimisation proximale afin de résoudre efficacement ces problèmes en dimension élevée (voir publications 4, 18). J'ai utilisé avec succès ces techniques dans un projet de collaboration internationale avec le laboratoire BRAMS à Montréal ainsi que dans la direction d'une partie d'un travail de doctorat en réalisant la détection de signatures motrices individuelles à partir de données de mouvement (voir publications 3, 5). Ces recherches allient fortement le versant mathématique de l'optimisation avec le versant applicatif de l'apprentissage automatique.

Ma contribution en optimisation continue à trouver plusieurs applications en recherche et industrie (voir projets PR2, PR4, PR7, PR8, PR12). Une solution performante pour le calcul en temps réel des rejets toxiques des mélanges pétroliers a fait l'objet d'un brevet international (voir publication 2). Il

s'agit de fabriquer des mélanges en respectant des normes environnementales de plus en plus restrictives.

La connaissance de la théorie des graphes m'a permis d'appliquer l'optimisation combinatoire dans plusieurs projets industriels (voir projets PR1, PR2, PR6, PR10, PR16). Un exemple est un problème de placement optimal dans un graphe afin de désengorger les centres-villes. Le test réalisé avec les hubs optimaux pour la ville de Rome a permis de diminuer de 15% les trajets des véhicules de service. Le calcul d'un parcours eulérien optimal (problème du postier chinois) dans un graphe orienté est un problème d'optimisation combinatoire NP-difficile contrairement au problème classique non-orienté. L'optimisation du ramassage de déchets porte-à-porte dans le cadre du projet européen Wasman a permis de réduire de 5% (d'un total de 1 500 000 km) les trajets. Ce même type de problème a été résolu pour la planification optimale de soins itinérants (Medtruck). Un autre projet utilise les modèles de diffusion dans un graphe pour évaluer l'influence dans un réseau social (Accrétion).

Apprentissage artificiel

Ma contribution dans le domaine de l'apprentissage artificiel a commencé durant mon deuxième contrat post-doctoral avec la société ELF Antar France en 1997. Le sujet était : Quelle confiance accorder à un modèle de réseau de neurones ? ». A cette époque les réseaux de neurones atteignaient leur maturité et les modèles à support vector montraient des caractéristiques prometteuses. C'était l'occasion de découvrir ce domaine passionnant fortement lié aux mathématiques de l'optimisation et plus en profondeur avec les théories de l'information et les statistiques.

La confiance est reliée avec la quantité de l'information fournie par une connaissance expert et / ou par les données. L'information fournie par les données est estimée par l'entropie de la densité de probabilité dont sont issues les données, avec la supposition sous-jacente (presque jamais vérifiée !) que l'échantillon fourni est représentatif de l'ensemble des données du phénomène étudié. Le calcul non paramétrique de la densité de probabilité passe par des fonctions noyaux qui emploient des distances. Ainsi, la question fondamentale sous-jacente était : quelle est la norme adéquate pour mesurer la distance entre les observations ? Les notions de distance et de similarité constituent un élément central dans l'estimation de l'information et donc de la confiance sur les résultats d'un modèle. Nous verrons par la suite que cette question est présente dans plusieurs travaux et en particulier elle constitue la problématique centrale d'une thèse de doctorat que j'ai co-dirigée (thèse T4).

Pendant ce même post-doctorat j'ai découvert les support vector machines, un modèle mathématique élégant et fortement lié à l'optimisation. Pour le problème de classement en deux classes séparables, on introduit la notion du gap. Lorsque le gap entre les classes est mesuré par la norme quadratique, la solution du problème de classement se réduit à la solution d'un problème d'optimisation quadratique sous contraintes linéaires. Les contraintes représentent le fait que les deux classes doivent être séparées. Ainsi, pour le problème de classement binaire, les support vectors sont géométriquement les observations les plus proches de la surface séparatrice optimale, celle qui maximise le gap de deux classes. Dans la réalité des calculs, les support vectors sont les observations dont les variables duales correspondantes valent 0. Encore la convexité !

Après cinq années enrichissantes de travail dans l'industrie j'ai voulu me consacrer plus à la recherche et l'enseignement. L'argument que j'ai présenté lors du jury de recrutement visait un ré-équilibrage entre les trois activités qui m'attiraient le plus : recherche, enseignement, applications. Le poste

d'enseignant-chercheur à l'Ecole des Mines d'Alès correspondait à mon profil. Depuis octobre 2000 j'exerce cette fonction au sein du LGI2P et maintenant dans l'UMR EuroMov DHM.

Depuis mon deuxième contrat post doctoral j'ai appris que la recherche d'information dans les données aboutit souvent à des modèles d'apprentissage. A leur tour, ces modèles s'écrivent en grande partie comme des problèmes d'optimisation continue, parfois convexe. Mais, ce chemin d'extraction d'information est plus long en réalité et l'appellation « data scientist » correspond mieux aux différents aspects de cette discipline ancienne pour la problématique et nouvelle par les technologies utilisées.

Avant d'arriver à l'apprentissage proprement dit, les données passent par plusieurs étapes de traitement. En premier lieu, il faut assurer l'extraction de variables significatives à partir des données de départ. Le traitement d'image et l'extraction d'informations est un exemple typique en la matière. Est-il facile de détecter l'œil, un visage, dans une image ? Est-ce qu'on trouve le même objet dans deux images ? C'est ce type de questions que j'ai été amené à traiter à travers la direction d'une thèse de doctorat (thèse T3). Trouver le même objet dans deux images est connu sous le nom « mise en correspondance ». Il existe des modèles mathématiques qui quantifient cette mise en correspondance de points d'intérêt dans l'image (coin, arête, …). Une idée intéressante pour augmenter la probabilité de correspondance est d'ajouter des objets tiers, A et B correspondent s'ils sont voisins avec C dans les deux images. Cette formulation peut être traduite en un problème d'optimisation quadratique avec Hessien non-symétrique. Un problème d'optimisation conçu sur mesure a permis d'accélérer d'un facteur important la mise en correspondance et a significativement amélioré sa qualité (thèse T3, voir publications 29, 30, 31, 32).

Ensuite, il est admis dans les règles de l'art de l'apprentissage qu'il faut partager l'ensemble des données au moins en deux sous-ensembles : apprentissage et test. Une mauvaise partition peut anéantir les qualités des algorithmes d'apprentissage. Les géants d'internet en font aujourd'hui les frais et sont accusés de « racisme » parce que leurs algorithmes ont appris avec des sous-ensembles de données qui ne couvrent pas uniformément le domaine étudié. Ainsi, on doit répondre à la question : Comment partager les données afin que chaque sous-ensemble représente « fidèlement » le tout ? Dans l'autre sens, lorsqu'on peut choisir la production des données, cette question est abordée par les techniques des plans d'expérience. La « fidélité » de la partition peut être quantifiée par la ressemblance des surfaces de densité calculées sur chacun des ensembles. Le calcul des densités étant assez long, d'autres caractéristiques de la densité de probabilité peuvent être utilisées comme les barycentres, les longueurs des axes propres des matrices de covariance, etc. Un critère de fidélité peut être calculé à partir de ces indices. Il s'agit ensuite de générer des sous-ensembles et de trouver une partition qui maximise le critère. C'est un problème d'optimisation combinatoire. Un premier logiciel qui utilise l'heuristique des algorithmes génétiques pour la recherche d'une « bonne » partition est produit pour l'industrie. Ensuite, nous avons utilisé l'heuristique tabou-search afin de trouver une partition (voir publication 37).

Ce n'est qu'après que viennent les algorithmes d'apprentissage proprement dit. Une hypothèse fondamentale de plusieurs méthodes d'apprentissage est : deux objets « similaires » doivent avoir des réponses similaires. Il est clair que cette hypothèse est réductrice et ne prend pas en compte les situations où un changement infime sur les variables explicatives peut induire un changement important de classe ou d'état de fonctionnement. Cette hypothèse est intégrée comme condition suffisante dans plusieurs méthodes d'apprentissage (réseaux de neurones profonds ou pas, support vector machines, arbres de décision, régression classiques ou lasso, …) par une condition

Lipschitzienne de loss-fonction. Elle assure que la distance entre deux réponses $f(x)$ et $f(y)$ est bornée par un facteur multipliant la distance entre x et y :

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

Le mot distance apparait deux fois dans la phrase (et la formule) précédente mais la condition ne dit pas quelle norme utiliser. Ce qui est un sujet central : comment mesurer la distance entre deux objets ? Et la similarité ? Et la dissimilarité ? Souvent ces notions sont confondues alors que la similarité ne vérifie pas l'inégalité triangulaire comme la distance. Aussi, connaitre la dissimilarité entre A et B et celle entre B et C ne donne aucune indication sur la dissimilarité entre A et C ! Le travail sur les mesures de similarité qui a commencé depuis mon deuxième post-doctorat a pris toute son ampleur avec l'encadrement majeur d'une thèse (livre 1, articles 7, 8, 15, 19, 20, 23, 24, 25, 26, 27, 28, 38, 40).
Le premier constat fut un très grand nombre de mesures utilisées partout avec des noms différents, « tout le monde y va de sa mesure ! » Après un long travail pour comparer toutes ces mesures, nous avons montré qu'une très grande partie des mesures de similarité, actuellement dans la littérature, sont des cas particuliers d'une formulation abstraite du '*ratio model*' introduit par Tversky :

$$sim_{RM}(u, v) = \frac{\Psi(\tilde{u}, \tilde{v})}{\alpha\ \Phi(\tilde{u}, \tilde{v}) + \beta\ \Phi(\tilde{v}, \tilde{u}) + \Psi(\tilde{u}, \tilde{v})}$$

Ici, la primitive $\Psi(\tilde{u}, \tilde{v})$ mesure la commonalité entre u et v, $\Phi(\tilde{u}, \tilde{v})$, $\Phi(\tilde{v}, \tilde{u})$ quantifie l'information dans u (v) qui n'est pas v (u). Les paramètres α et β sont ajustables. In fine, trois primitives et deux paramètres. Ce cadre élégant a été traduit en un outil informatique pour choisir les primitives et ajuster les paramètres.

Depuis deux ans, je participe aux activités de recherche du laboratoire EuroMov à Montpellier qui se consacre à l'étude du mouvement humain. A partir de janvier 2021, je ferai partie de l'UMR EuroMov Digital Health in Motion – Université de Montpellier et IMT Mines Alès. C'est le cadre applicatif idéal pour toute la panoplie de concepts, méthodes et outils en ma possession et un champ de questions et de problèmes qui nécessite une montée en compétences fondamentales sur des nouvelles approches et techniques. Le mouvement humain mobilise plusieurs mécanismes sensoriels et moteurs. La défaillance d'un de ces mécanismes, causée par exemple par des maladies neuro-dégénératives, engendre des perturbations du mouvement, perturbations qui peuvent parfois être invisibles à l'œil nu. Ainsi, le mouvement peut fournir des indicateurs précoces de ces maladies. Encore, faut-il être en mesure d'extraire cette information.

De grandes quantités de données sont enregistrées, de typologies et natures différentes impliquant la capture de mouvement, électroencéphalogrammes, électrocardiogrammes, etc. Une caractéristique fondamentale de ces données est un nombre d'observations bien inférieur à la dimension des données. Ceci est dû au fait que les observables sont des humains et le nombre de participants dans une expérience peut rarement dépasser quelques dizaines alors que les données mentionnées, générées à des fréquences élevées, peuvent atteindre des dimensions de dizaines de milliers. Tous ces cas d'application subissent la « malédiction » de la dimension. Heureusement, ces données présentent une redondance importante entre les variables physiologiques. Cette redondance permet de réduire efficacement la dimension des données pour obtenir un petit nombre de variables explicatives.

Quand on regarde de près plusieurs méthodes d'apprentissage, il s'agit souvent de transformer les données initiales de dimension d en les projetant dans un espace (ou une suite d'espaces) de dimension $d' >> d$ avant de calculer une réponse. Ce procédé complique, voire rend impossible, la capacité d'expliquer la réponse. Ce qui donne à plusieurs méthodes un aspect de « boite noire ». Or, les données physiologiques mentionnées sont déjà de dimension élevée. Ainsi j'ai choisi d'orienter mon activité de recherche vers les méthodes de réduction de dimension en cherchant un petit nombre de variables explicatives parmi les variables physiologiques d'origine. Il est ainsi plus aisé pour les spécialistes d'interpréter les résultats. Au-delà des méthodes classiques telles que lasso, je travaille sur des méthodes d'optimisation proximale (voir publications 4, 18). Nous avons utilisé ces méthodes pour la recherche de signatures motrices à partir de données de Motion Capture (voir thèse T8 et publication 3). Cette même méthode a réussi à détecter les profils de musiciens dans le cadre du projet de recherche BRAMS (PR18).

Mes travaux de recherche m'ont amené à encadrer, ou contribuer scientifiquement à :

- Dix thèses de doctorat : sept avec taux d'encadrement importants et trois avec taux d'encadrement plus faibles sur des contributions méthodologiques et applicatives ;
- Trois post-doctorats ;

Une projection de ces travaux sur les axes principaux de mes compétences donne :

Apprentissage artificiel

Thèses : T3, T4, T5, T6, T7, T8, T9, T10.

Post doctorats : P1, P2, P3.

Principales publications : 1, 3, 4, 5, 7, 8, 9, 10, 12, 15, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42.

Principaux projets : PR3, PR5, PR9, PR10, PR11, PR13, PR14, PR15, PR16, PR18, PR19.

Optimisation

Thèses : T1, T2, T3

Principales publications : 2, 4, 6, 11, 13, 14, 18, 21, 22, 29, 30, 31, 32, 33, 34, 35, 37.

Principaux projets : PR4, PR7, PR8, PR12.

Graphes, autres

Principales publications : 10, 12, 13, 14, 43.

Principaux projets : PR1, PR2, PR6, PR17.

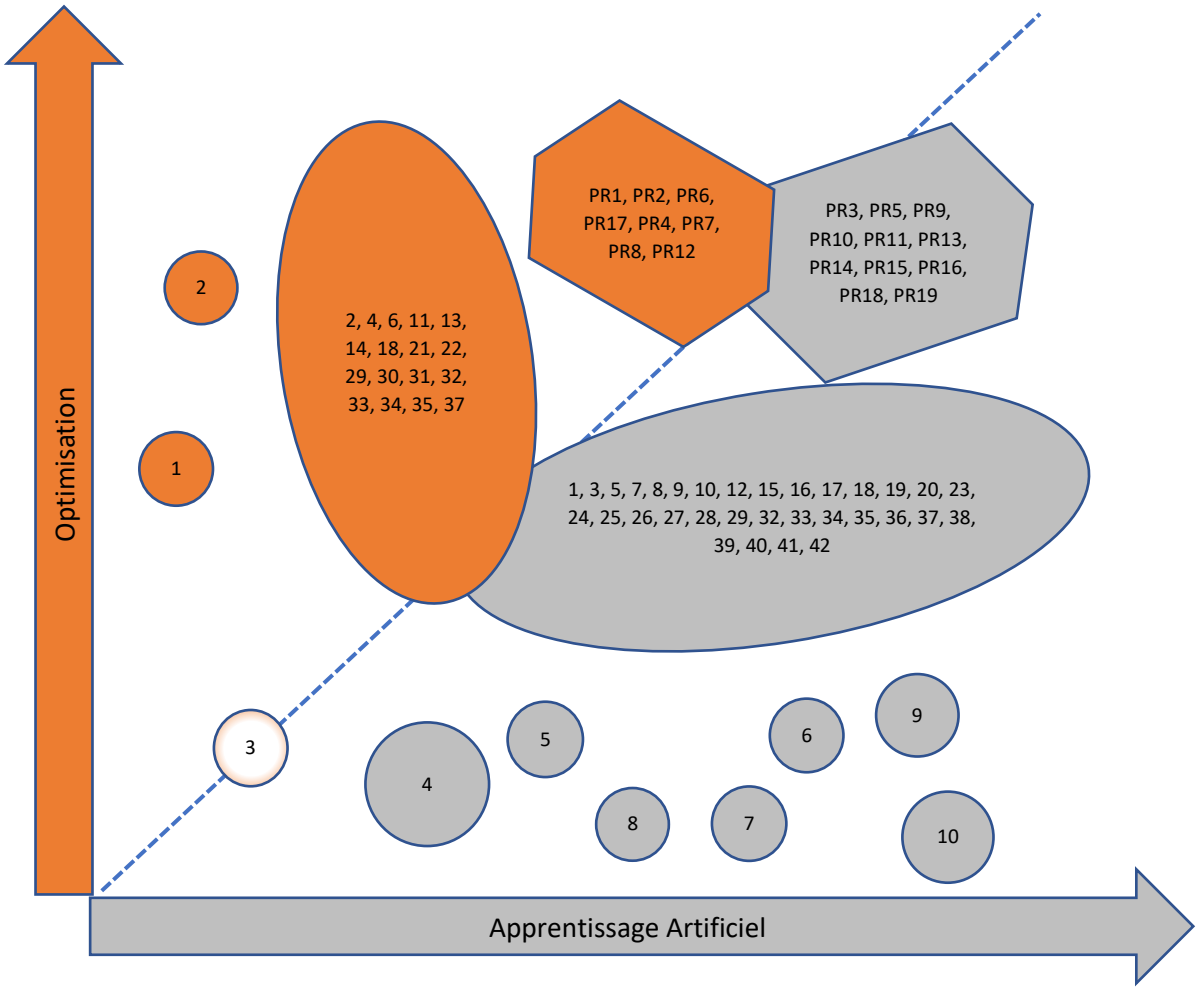

Figure 2. Une représentation schématique de mes contributions en direction de recherche (cercles), publications (ellipses) et projets industriels (hexagones).

2. Encadrement de doctorants

T1 – Vitaliy Feoktistov – Contribution à l'Évolution Différentielle

Ecole doctorale : Ecole Nationale Supérieure des Mines de Paris
Spécialité : Informatique temps réel, Automatique et Robotique
Préparée au : LGI2P, IMT – Mines Alès
Soutenue en décembre 2004
Jury de thèse :
M. Alexandre Dolgui – Professeur ENMSE – Rapporteur
M. Michel Habib – Professeur LIRMM – Université Montpellier 2, Rapporteur
M. Yves ROUCHALEAU – Professeur, CMA – ENSMP – Directeur de Thèse
M. Jacky MONTMAIN – HDR EMA-CEA – Directeur de thèse
M. David PEARSON – Professeur, Université Jean Monnet Saint Etienne – Examinateur
M. Stefan JANAQI – Docteur – LGI2P – EMA – Encadrant de thèse

Contribution à l'encadrement : 50%.

Situation actuelle : Optimization, HPC and Deep Learning LIMAGRAIN agro-alimentaire, Clermont-Ferrand.

Mots clés : Optimisation Mathématique, Evolution, Métaheuristique, Stochastique, Algorithme.

Résumé
Le domaine des algorithmes évolutionnaires a connu un grand développement ces dernières années. L'évolution différentielle (ED) est l'un de ces algorithmes. À l'origine, l'ED était conçue pour les problèmes d'optimisation continus et sans contraintes. Ses extensions actuelles peuvent traiter des problèmes en variables mixtes sous contraintes non linéaires. Depuis, l'ED est devenue une méthode efficace pour une grande quantité de problèmes réels. L'objectif principal de cette thèse était de proposer une analyse exhaustive de cette méthode : sur un plan scientifique, de la situer par rapport aux méthodes plus classiques concurrentes et sur un plan technique, d'améliorer son taux de convergence et d'augmenter sa capacité à trouver l'optimum global. Dans ce dessein, nous décelons dans l'algorithme trois niveaux hiérarchiques à améliorer. Au premier niveau le comportement d'un individu a été étudié. Une formule universelle de la variation d'individu est mise au point : elle permet d'envisager la création d'un grand nombre de stratégies d'exploration et d'exploitation de l'espace de recherche. Initialement, quatre stratégies sont proposées. Une généralisation de l'algorithme est réalisée en introduisant l'ED transversale. Au deuxième niveau, la population entière est considérée. Afin d'accélérer la convergence, le principe de la sélection énergétique a été mis au point et testé, il permet de retenir les individus réputés les meilleurs. Une interaction de l'algorithme principal avec une méthode de régression externe est réalisée pour compléter la sélection des individus les plus intéressants et augmenter in fine le taux de convergence. Concrètement, ce troisième niveau implémente l'interaction entre l'ED et SVM. L'idée centrale de cette approche est de trouver un bon compromis entre l'exploration de l'espace de recherche et l'exploitation de l'information locale envoyée par chaque point généré. Dans ce but, un indicateur de diversité est introduit pour quantifier

la capacité de l'exploration d'une stratégie utilisée. Cet indicateur utilise des caractéristiques géométriques du nuage de points en évolution telles que le diamètre, la dimension de la variété linéaire les contenant, la longueur des axes de l'ellipsoïde défini par le Hessian de l'approximation de degré deux des couples $(x, f(x))$ du nuage. Ces caractéristiques géométriques permettent de définir des stratégies d'évolution efficaces et des critères d'arrêt robustes. Du point de vue complexité calculatoire, l'évolution différentielle est une méthode nativement parallélisable. Elle a montré ces capacités de trouver des optimums satisfaisant sur de larges benchmarks de problèmes académiques ou industriels (articles 35, 36, 37).

T2 – Jorge Aguilera – Robustesse et visualisation de production de mélanges

Ecole doctorale : MSTII - mathématiques, sciences et technologies de l'information, informatique (Grenoble) – en partenariat avec Institut Fourier (Grenoble)
Préparée au : LGI2P – Ecole des Mines d'Alès
Spécialité : Mathématiques appliquées
Soutenue en octobre 2011
Jury de thèse :
Mme. Nadia Maïzi-Ménard – Professeur Ecole des Mines de Paris – Présidente du jury.
M. Ridha A. Mahjoub – Professeur des Universités, Université Paris Dauphiné – Rapporteur
M. Alexandre Dolgui – Professeur ENMSE – Rapporteur
M. Michel Mollard – HDR EMA-CEA – Directeur de thèse
M. Stefan JANAQI – Docteur – LGI2P – EMA – Encadrant de thèse
Mme. Meriam Chèbre – TOTAL – Invitée

Contribution à l'encadrement : 70%.

Situation actuelle : Banque du Mexique, Yucatan. Chargé de planification et implémentation de projets d'optimisation.

Mots clés : Problème de mélange, Optimisation Robuste, Polyèdres, Infaisabilité, Visualisation.

Résumé
Le procédé de fabrication de mélanges (PM) consiste à déterminer les proportions optimales à mélanger un ensemble de composants de façon que le produit obtenu satisfasse un ensemble de spécifications sur leurs propriétés. Deux caractéristiques importantes du problème de mélange sont d'une part, les bornes non violables (car réglementaires) sur les propriétés du mélange et d'autre part, l'incertitude inhérente au procédé. Contrôler en temps réel ces deux caractéristiques antagonistes pour de longues séquences de mélanges est un défi mathématique et calculatoire. Dans ce travail, on propose une méthode pour la production de mélanges robustes en temps réel qui minimise le coût de la recette et la sur-qualité du mélange. La notion d'optimisation robuste, en particulier en programmation linéaire a été formalisé par Ben-Tal et Nemirovskii. Leur analyse est basée sur la norme quadratique de la variation des contraintes. Ainsi, pour se prémunir de l'incertitude des données, on vise des solutions faisables dont les contraintes linéaires de base sont augmentées par la norme de variation des contraintes. Le problème obtenu devient un problème d'optimisation convexe. Afin de satisfaire les contraintes temps-réel (le solveur est appelé toutes les 4s en moyenne), nous avons reformulé le cadre de robustesse en employant la norme L_1 et sa norme duale L_∞. Les contraintes du problème robuste sont moins strictes que pour la norme L_2, en revanche, ce problème reste linéaire et peut utiliser tout le trésor des solveurs de programmes linéaires. La méthode est basée sur l'hypothèse que les lois des mélanges sont linéaires. Un aspect important de ce travail est

la visualisation des polytopes de mélange ou plutôt de coupes d'intérêt de ces polytopes. Cette visualisation permet de rendre plus concret pour les opérateurs le process de mélange. Aussi, cela permet de caractériser l'infaisabilité du PM et d'analyser la modification des bornes sur les composants pour guider le procédé vers le ``meilleur`` mélange robuste. On propose un ensemble d'indicateurs et de visualisations en vue d'offrir une aide à la décision (articles 27, 28).

T3 – Désiré Sidibé – Une technique de relaxation pour la mise en correspondance d'images : application à la reconnaissance d'objets et au suivi du visage

Ecole doctorale : Information, Structure, Systèmes – Université II Montpellier
Préparée au : LGI2P – Ecole des Mines d'Alès
Spécialité : Informatique
Soutenue en décembre 2007
Jury de thèse :
Mme. Christine Fernandez-Maloigne – Professeur, Université de Poitiers – Rapporteur.
M. Frédéric Jurie – Professeur, Université Caen.
M. Jean-Claude Bajard – Professeur, Université Montpellier II, Directeur de thèse.
Mme Valérie Gouet – Enseignant-Chercheur, CNAM Paris – Examinateur.
M. Philippe Montésinos – Enseignant-Chercheur, Ecole des Mines d'Alès – Encadrant de proximité.
M. Stefan Janaqi – Enseignant-Chercheur, Ecole des Mines d'Alès – Encadrant de proximité.
M. Réné Zapata, Professeur, Université de Montpellier II, Examinateur.

Contribution à l'encadrement : 50%.

Situation actuelle : Maître de conférences à l'Université de Bourgogne.

Mots clés : Mise en correspondance d'images, reconnaissance d'objets, relaxation, détection de la peau, détection du visage, suivi du visage.

Résumé : Le principal intérêt de l'utilisation des invariants locaux pour la mise en correspondance de différentes vues d'une même scène est le caractère local qui les rend robustes aux occultations et aux changements de point de vue et d'échelle. Une autre caractéristique importante est le calcul efficace de cette information locale. Néanmoins, cette localité limite le pouvoir discriminant des descripteurs locaux qui échouent dans les cas difficiles où l'ambiguïté est élevée. Dans une première partie, nous proposons une méthode de mise en correspondance basée sur la relaxation qui prend en compte une information plus globale, dite contextuelle, afin de garantir des résultats corrects même dans les cas les plus difficiles. Nous présentons une application dans le cadre de la reconnaissance d'objets dans des scènes complexes. Etant donné un objet A et son correspondant recherché A' dans deux images, la notion de contexte est matérialisée par la recherche simultané de voisins B et B' de A et A' respectivement. Ainsi, on ne cherche pas seulement la correspondance entre A et A' mais aussi celle des voisins B et B'. Cette recherche se fait dans un cadre probabiliste qui cherche parmi tous les objets possibles celui qui maximise la probabilité de correspondance. Le contexte mentionné peut se modéliser par une matrice de voisinage non-symétrique. Ceci transforme le problème de mise en correspondance en un problème d'optimisation quadratique sur le simplexe de probabilités avec un Hessian non-symétrique. Un solveur d'optimisation sur mesure (puisque tous les solveurs classiques de problèmes quadratiques sous contraintes linéaires ont un Hessian symétrique et défini positif pour assurer la convexité du problème) a permis de trouver des résultats dépassant par le taux de réponse correcte et en vitesse les méthodes concurrentes. Dans une seconde partie, nous abordons le problème de la détection et du suivi du visage dans une séquence d'image. Nous proposons une

méthode simple et efficace pour la détection du visage dans une image couleur. Il s'agit d'exploiter la particularité biométrique et géométrique des yeux pour calculer rapidement s'il s'agit d'un visage ou non. Ce critère est couplé avec l'algorithme de mise en correspondance pour suivre efficacement le visage dans une séquence d'images (articles 31, 32, 33).

T4 – Sébastien Harispe – Knowledge-based Semantic Measures: From Theory to Applications

Ecole doctorale : Information, Structure, Systèmes – Université II, Montpellier
Préparée au : LGI2P – Ecole des Mines d'Alès
Spécialité : Informatique
Soutenue en avril 2014
Jury de thèse :
M. Jérôme Euzenat – Directeur de recherche, INRIA Grenoble Rhône-Alpes - Rapporteur
Mme. Pascale Kuntz-Cosperec – Professeur des Universités, Ecole Polytechnique de l'Université de Nantes – Rapporteur
M. Amedeo Napoli – Directeur de Recherche CNRS, LORIA, Nancy – Président du Jury
M. Jacky Montmain – Professeur, IMT – Mines Alès, Directeur de thèse
Mme Isabelle Mougenot – Maître de Conférences, LIRMM, Université Montpellier II – Examinateur
M. David Sanchez – PhD, Universitat Rovira i Virgili, Tarragona (ES) – Examinateur
M. Stefan Janaqi, Enseignant-Chercheur, IMT Mines Alès, Encadrant de proximité
Mme Sylvie Ranwez, Enseignant-Chercheur, IMT Mines-Alès, Encadrant de proximité

Contribution à l'encadrement : 40%.

Situation actuelle : Enseignant-Chercheur à l'IMT Mines Alès.

Mots clés : Mesures sémantiques, ontologie, algorithme, similarité sémantique, web sémantique, ontologies, gestion de connaissances.

Résumé
Les notions de proximité, de distance et de similarités sémantiques sont depuis longtemps jugées essentielles dans l'élaboration de nombreux processus cognitifs et revêtent donc un intérêt majeur pour les communautés intéressées au développement d'intelligences artificielles. Cette thèse s'intéresse aux différentes mesures sémantiques permettant de comparer des unités lexicales, des concepts ou des instances par l'analyse de corpus de textes ou de représentations de connaissance (e.g. ontologies). Encouragées par l'essor des technologies liées à l'Ingénierie des Connaissances et au Web sémantique, ces mesures suscitent de plus en plus d'intérêt à la fois dans le monde académique et industriel. Ce manuscrit débute par un vaste état de l'art qui met en regard des travaux publiés dans différentes communautés et souligne l'aspect interdisciplinaire et la diversité des recherches actuelles dans ce domaine. Cela nous a permis, sous l'apparente hétérogénéité des mesures existantes, de distinguer certaines propriétés communes et de présenter une classification générale des approches proposées. Par la suite, ces travaux se concentrent sur les mesures qui s'appuient sur une structuration de la connaissance sous forme de graphes sémantiques, e.g. graphes RDF(S). Nous montrons que ces mesures reposent sur un ensemble réduit de primitives abstraites, et que la plupart d'entre elles, bien que définies indépendamment dans la littérature, ne sont que des expressions particulières de mesures paramétriques génériques. Ce résultat nous a conduits à définir un cadre théorique unificateur pour les mesures sémantiques. Il permet notamment : (i) d'exprimer de nouvelles mesures, (ii) d'étudier les propriétés théoriques des mesures et (iii) d'orienter l'utilisateur dans le choix d'une mesure adaptée à sa problématique. Les premiers cas concrets d'utilisation de ce

cadre démontrent son intérêt en soulignant notamment qu'il permet l'analyse théorique et empirique des mesures avec un degré de détail particulièrement fin, jamais atteint jusque-là. Plus généralement, ce cadre théorique permet de poser un regard neuf sur ce domaine et ouvre de nombreuses perspectives prometteuses pour l'analyse des mesures sémantiques. Le domaine des mesures sémantiques souffre d'un réel manque d'outils logiciels génériques et performants ce qui complique à la fois l'étude et l'utilisation de ces mesures. En réponse à ce manque, nous avons développé la Semantic Measures Library (SML), une librairie logicielle dédiée au calcul et à l'analyse des mesures sémantiques. Elle permet d'utiliser des centaines de mesures issues à la fois de la littérature et des fonctions paramétriques étudiées dans le cadre unificateur introduit. Celles-ci peuvent être analysées et comparées à l'aide des différentes fonctionnalités proposées par la librairie. La SML s'accompagne d'une large documentation, d'outils logiciels permettant son utilisation par des non informaticiens, d'une liste de diffusion, et de façon plus large, se propose de fédérer les différentes communautés du domaine afin de créer une synergie interdisciplinaire autour la notion de mesures sémantiques : http://www.semantic-measures-library.org. Cette étude a également conduit à différentes contributions algorithmiques et théoriques, dont (i) la définition d'une méthode innovante pour la comparaison d'instances définies dans un graphe sémantique – nous montrons son intérêt pour la mise en place de système de recommandation à base de contenu, (ii) une nouvelle approche pour comparer des concepts représentés dans des taxonomies chevauchantes, (iii) des optimisations algorithmiques pour le calcul de certaines mesures sémantiques, et (iv) une technique d'apprentissage semi-supervisée permettant de cibler les mesures sémantiques adaptées à un contexte applicatif particulier en prenant en compte l'incertitude associée au jeu de test utilisé. Ces travaux ont été validés par plusieurs publications et communications nationales et internationales.

T5 – Grégoire Vergotte – Adaptability and adaptation to a sensorimotor task: from functional significance of fractal properties to brain networks dynamics

Thèse de l'Université de Montpellier
Ecole doctorale : Sciences du Mouvement Humain
Spécialité : Informatique
Soutenue en novembre 2018
Jury de thèse :
M. Frédéric Dehais - Professeur ISAE Supaero - Rapporteur
M. Giovanni de Marco - Professeur Université Paris Nanterre - Rapporteur
M. Jean-Jacques Temprado – Professeur - Université de Marseille – Président du Jury.
Mme. Emmanuelle le Bars – Docteur CHU Montpellier - Examinateur
M. Stéphane Perrey - Professeur Staps-EuroMov- Université Montpellier – Directeur de thèse
Mme. Kjerstin Torre - Maître de conférences-Université de Montpellier – Co-encadrant

Contribution à l'encadrement : 20%
Situation actuelle : Post Doctorat
Mots clés : Dégénérescence, adaptabilité, fractales, réseau, SPIR, adaptation.

Résumé
L'étude des propriétés fractales des séries biologiques fait l'objet d'un intérêt croissant. Néanmoins la littérature met en évidence une ambiguïté quant à l'explication causale de la présence de ces séries temporelles ne permettant pas de distinguer entre l'adaptation effective réalisée par un sujet ou ses capacités d'adaptabilité globales. La présente thèse a pour objectif de décorréler ces deux notions,

notamment en liant le niveau comportemental au niveau cérébral. Notre première étude a permis de mettre en évidence que les propriétés mono-fractales pourraient refléter l'adaptabilité des sujets tandis que les propriétés multifractales seraient liées à l'adaptation effective réalisée au cours de la tâche. La seconde étude a mis en évidence une corrélation entre les propriétés multifractales et le nombre de réseaux cérébraux mis en œuvre au cours de la tâche, reflétant l'adaptation effective aux contraintes expérimentales imposées. Les résultats de ces travaux de thèse nous ont permis de mieux comprendre la signification fonctionnelle des analyses fractales en termes d'adaptation effective et d'adaptabilité. Considérant une matrice de similarité dont les éléments sont les corrélations entre les observations nous avons réussi à détecter des communautés de comportement par une méthode de maximisation de la modularité. C'est une heuristique qui repose sur la décomposition spectrale d'une matrice de modularité, calculée à partir de la matrice de similarité. Cette matrice de modularité joue dans la détection de communautés un rôle similaire que la Laplacienne dans le problème de partition de graphes.

T6 - Robert QUACH - Identification d'un modèle flou appliqué à un problème de classification

Thèse de l'Université Jean Monnet de Saint-Etienne
Préparée au LGI2P – IMT Mines Alès
Ecole doctorale : Sciences pour l'Ingénieur
Spécialité : Informatique
Soutenue en septembre 2002
Jury de thèse :
M. José Ragot - Professeur Université de Nancy - Rapporteur
M. Jean-François Santucci - Professeur Université de Corse - Rapporteur
M. Eric Matzner-Lober – Maître de Conférences - Université de Rennes II – Examinateur
M. Jacky Montmain - Professeur Ecole des Mines d'Alès - Examinateur
M. David Pearson - Professeur Université Jean Monet St Etienne – Directeur de thèse
M. Gérard Dray - Maître assistant - Ecole des Mines d'Alès – Co-encadrant

Contribution à l'encadrement : 10%
Situation actuelle : Ingénieur de Recherche CEA
Mots clés : Classification, Regroupement par Soustraction, Logique Floue, Courbe

Résumé
Dans cette thèse, nous cherchons à identifier le modèle d'un système complexe. Après avoir décrit le type de modèle que nous retenons, c'est-à-dire le Système d'Inférence Flou (SIF), nous décrivons l'approche qui permet de l'identifier. Cette approche consiste à effectuer d'abord un regroupement flou, puis les groupements flous sont projetés sur les axes pour obtenir le SIF. La méthode de regroupement que nous utilisons dans notre étude est le Regroupement par Soustraction (RS). Notre première contribution se résume par l'extension de cette méthode d'abord pour diminuer les problèmes liés à la dimension des données (MRS), ensuite pour faciliter la détermination des valeurs de paramètre et enfin pour identifier des groupements de taille différente afin d'augmenter la parcimonie du modèle flou (RSE). Nous appliquons ensuite l'indentification du modèle à un problème de classification de courbes fonctionnelles. L'idée principale consiste à considérer la séquence d'observations des courbes comme une seule entité, et non comme des attributs individuels. Pour identifier le modèle classifieur flou, nous proposons d'utiliser tout d'abord des métriques plus adéquates permettant de tenir compte des variations de courbe. Puis, nous proposons une nouvelle approche non-supervisée dans le but de réduire la description des courbes tout en gardant leur forme

évolutive. Enfin, nous utilisons l'approche de changement d'espace en b-spline en espérant que cette transformation permette d'ajouter une information substantielle pour le processus de la classification. La contribution générale de toute notre étude est que nous avons proposé des méthodes qui permettent d'obtenir des modèles plus parcimonieux, et donc de généraliser la connaissance.

T7 – Pascale Montréer. Reliability improvement of Odour Detection

Thèse de l'Université UNimes à Nîmes
Préparée au LGI2P et LGEI – IMT Mines Alès
Ecole doctorale : Risques et Société
Spécialité : Biostatistique
Soutenue en novembre 2019

Jury de thèse :
El Mostafa QANNARI, Professeur, ONIRIS Nantes – Rapporteur
Pierre LE CLOIREC, Professeur, Ecole Nationale Supérieure de Chimie de Rennes – Rapporteur
Anne BERGERET, Professeur, IMT Mines Alès – Examinateur
Marie-France FALZON, Dr, Responsable du laboratoire d'analyses physico-chimiques, HUTCHINSON, Chalette-sur-Loing – Examinateur
Mathilde CHAIGNAUD, Dr, Chargée d'études, OLENTICA, Alès – Examinateur
Jean-Louis FANLO, Professeur, IMT Mines Alès – Directeur de thèse
Stefan JANAQI, Enseignant-chercheur, IMT Mines Alès – Co-Encadrant
Stéphane CARIOU, Enseignant-chercheur, IMT Mines Alès – Co-Encadrant

Contribution à l'encadrement : 40%.
Situation actuelle : Biostatisticienne chez IT & M STATS.
Mots clés : Odour Detection Thresholds (ODT), Data mining, Reliability, Completeness, Uncertainty.

Résumé
Dans le milieu industriel, les matériaux générant une odeur désagréable représentent une problématique majeure. En effet, l'odeur fait très souvent partie des critères de sélection, d'achat et d'utilisation d'un produit par le consommateur. Si un matériau a une odeur désagréable, il risque d'être rejeté par le consommateur qui considérera sa qualité comme mauvaise ou altérée. Améliorer la qualité odorante d'un matériau constitue donc un enjeu industriel et économique important.
Dans ce contexte, le travail réalisé dans le cadre de cette thèse consiste à développer un protocole permettant d'identifier le ou les composés chimiques responsables de l'odeur désagréable d'un matériau. Ce protocole est développé sous forme d'un outil automatisé combinant une succession de techniques statistiques. L'un des piliers de ce travail est la recherche de corrélations entre la composition de la matrice gazeuse émise par le matériau (mesures physico-chimiques) et l'odeur associée à cette matrice (mesures olfactométrique). Pour atteindre cet objectif, un important travail d'investigation sur les seuils de perception olfactive (fiabilité et complétude des données) a été réalisé. Les données collectées sur les seuils de perception représentent de fortes variations selon les sources. Les sources sont de natures différentes : sources bibliographiques dont les résultats sont difficilement vérifiables ; sources d'essais en interne dont on maitrise toute la chaine de production des données. Cet élément est pris en compte dans nos modèles de prédiction en modulant les poids des observations. Ces seuils servent ensuite d'unité pour quantifier le niveau de l'odeur. Nous avons mis au point une prédiction des seuils à partir des données à base de régression SVM. Un compromis délicat entre la précision et la robustesse de prédiction est réalisé en utilisant des noyaux de faible

degré. Un programme informatique réalisant toutes les étapes de traitement de données jusqu'à la prédiction tourne actuellement en conditions industrielles.

T8 (en cours) Alexandre Coste. Les signatures motrices individuelles : de la perception humaine aux applications biométriques

Thèse de l'Université UM de Montpellier
Préparée au laboratoire EuroMov
Ecole doctorale : MSH

Directeur de thèse : Ludovic Marin, quotité de temps : 40 %, grade MCF HDR.
Co-directeur : Benoît Bardy, quotité de temps : 30%, grade PREX
Co-encadrant : Stefan Janaqi, quotité de temps : 30 %, grade MCF
Situation actuelle : Doctorat en cours (soutenance hiver 2020-2021)

Mots clés : Signatures motrices ; Similarité ; Mouvement biologique ; Perception de l'identité ; Biométrie.

Résumé
Le mouvement, composant primordial de notre existence, est utilisé quotidiennement pour accomplir des tâches des plus simples aux plus complexes, ou encore pour communiquer. Que cela soit intentionnel ou non, nos mouvements signent nos différences inter-individuelles (e.g., identité) mais aussi intra-individuelles liées aux états émotionnels, aux intentions, etc. Notre système visuel semble finement réglé pour percevoir et interpréter les informations contenues dans le mouvement biologique. Ce travail doctoral vise ainsi à mieux comprendre notre sensibilité au mouvement biologique, en particulier, celle liée à la perception de l'identité. Dans une série d'expérimentations, nous montrons : i) l'existence d'une signature motrice spécifique à chaque individu – une sorte d'empreinte cinématique – ayant pour caractéristiques principales une invariance temporelle et un caractère distinctif avec les autres signatures ; ii) une grande sensibilité des observateurs humains aux signatures motrices individuelles malgré les nombreuses sources de variations intra-individuelles telles que le changement de point de vue, la variabilité inter-essais iii) que la similarité cinématique entre les signatures est dramatiquement génératrice de confusion d'identité. Puisque ce problème de similarité n'est pas limité au système visuel humain mais se retrouve également dans la plupart des systèmes biométriques actuels, l'étude des signatures motrices individuelles se révèle une voie très prometteuse pour l'amélioration des taux de reconnaissance chez l'Homme et la machine (vision par ordinateur). Une caractéristique importante de ces données est un faible nombre d'observations en très grande dimension. Aussi, il y une très forte redondance dans les données par suite de la coordination des mouvements des parties du corps humain. Nous avons utilisé des méthodes de sparse learning pour diminuer la dimension et réaliser une identification des individus à partir d'enregistrement d'improvisations avec un taux d'erreur très faible. Un très petit nombre de variables explicatives réalisant cette identification constitue la signature motrice. La quantification de l'apport informationnel des différents points du corps a montré que le haut du corps réalise à lui seul une identification avec moins de 20% d'erreur.

T9 (en cours) Andrii Smikovski. Emotional interpersonal motor synchronisation.

Mots-clés de la thèse : Synchronisation de groupe, échelles temporelles multiples, expériences humaines, sciences des données.

Directeur de thèse : Benoit Bardy, quotité de temps : 50 %, grade PREX.
Co-encadrant : Stefan Janaqi, quotité de temps : 50 %, grade MCF.
Situation actuelle : Doctorat en cours

Résumé du projet de thèse : La synchronisation perceptuo-motrice de groupe est une caractéristique essentielle des activités humaines, par exemple lors des applaudissements du public, de la marche dans une foule, la musique, le sport et la danse. La synchronisation de groupe implique une intention partagée et une interaction perceptive, mais dépend également de la manière dont les signatures motrices individuelles sont assemblées pour former une signature motrice spécifique de groupe. Actuellement, la recherche est active dans le domaine du affective-computing afin d'identifier l'émotion des individus (Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.) (2015). The Oxford handbook of affective computing. Oxford, New York: Oxford University Press). L'information affective est extraite de marqueurs unimodaux tels que le mouvement, le regard, l'activation de muscles du visage, le rythme cardiaque, le rythme de respiration, etc. Tous ces marqueurs sont individuels et l'influence sociale sur les émotions est limitée voir absente dans les études (Fujiwara, K., & Daibo, I. (2018). Affect as an antecedent of synchrony: A spectrum analysis with wavelet transform. Quarterly Journal of Experimental Psychology, 71(12), 2520–2530.). Un de nos objectifs est de s'approcher de situations réelles où l'émotion est détectable dans le comportement de groupe. Notre approche vise l'utilisation de signatures motrices et leur quantification lorsqu'un individu est seul ou en groupe. Plus particulièrement, on cherche à mettre en évidence les effets des émotions de valences différentes (positive vs. neutres vs. négatives) qui sont induites expérimentalement. Au-delà de l'intérêt de répondre sur l'effet des émotions positives sur le degré de synchronisation (Paxton, A., & Dale, R. (2013). Argument disrupts interpersonal synchrony. Quarterly Journal of Experimental Psychology (2006), 66(11), 2092–2102), nous envisageons de chercher et quantifier des métriques qui caractérisent les états d'émotions positives, négatives et neutres.

T10 (en cours) Mélodie Sannier. Marcher à la maison Walk@Home.

Mots-clés de la thèse : Locomotion, Marche, Bien-être, Domicile, Santé, Science de données

Directeur de thèse : Benoit Bardy, quotité de temps : 50 %, grade PREX.
Co-directeur : Gérard Dray, quotité de temps : 25 %, grade PR.
Co-encadrant : Stefan Janaqi, quotité de temps : 25 %, grade MCF.
Situation actuelle : Doctorat en cours

Résumé du projet de thèse : Le projet de thèse Walk@Home a pour objectif de déterminer les marqueurs locomoteurs du bien-être à domicile. En effet, la façon dont l'Homme se déplace apporte de précieuses informations sur son bien-être physique et mental. C'est donc à travers un projet pluridisciplinaire liant sciences du mouvement humain, neurosciences et science de données que cette thèse va étudier les comportements locomoteurs de sujets sains dans un milieu non-contrôlé. Les données sont recueillies en continue dans un appartement connecté mis en place par le consortium HUman at home projecT. Des millions de données en provenance du sol connecté seront analysées afin d'approfondir les connaissances sur la locomotion humaine, mais également sur le bien-être à domicile. En analysant ces données on cherchera à comprendre l'occupation de l'espace, les zones « marchables » et les zones marchées, les vitesses de mouvement à l'intérieur, l'évolution de ces paramètres dans le temps (étude longitudinale), …

Nous nous intéressons à l'impact des attracteurs / répulseurs sur la marche. Quelles sont les caractéristiques observables de la marche qui changent en présence de perturbations extérieures ? L'identification des situations : marche, station, chute, … fera l'objet de nos recherches car leur application pour les populations vieillissantes est importante. L'extraction de ces informations à partir de ces données réelles est un vrai défi à cause de la grande taille des secteurs de base et de l'arrivée intermittente des activations. La détection d'une approximation de la trajectoire d'un individu passe par des notions de continuité spatio-temporelle qui doivent être adaptés pour les données en cours en prenant en compte des paramètres physiologiques tels que la longueur d'un pas, le temps de pose d'un pied, etc.

C'est donc à l'aide de l'étude des métriques de la marche (le pas, des trajectoires, de la vitesse de déplacement) que nous allons déterminer ces marqueurs de bien-être. A terme, les résultats de cette recherche permettront l'optimisation de l'espace architectural.

Encadrement de stages de Post Doc, DEA / Master 2

P1 – Mingyuan Jiu – **Post Doctorat Total Group in collaboration with ENS Lyon. Multiclass SVM with graph path coding regularization for face classification.**

Objective: Static and dynamic selection of subsets of observations and/or features in the frame of big data and active learning. Application to blending laws in petroleum industry.
Keywords: Multiclass SVM, graph path penalty, graphical Lasso, classification.

Abstract. We consider the problem of learning graphs in a sparse multiclass support vector machines framework. For such a problem, sparse graph penalty is useful to select the significant features and interpret the results. Classical L_1-norm learns a sparse solution without considering the structure between the features. In this paper, a structural knowledge is encoded as directed acyclic graph and a graph path penalty is incorporated to multiclass SVM. The learned classifiers not only improve the performance, but also help in the interpretation of the learned features. The performance of the proposed method highly depends on an initialization graph. Two generic ways to initialize the graph between the features are considered: one is built from similarities while the other one uses Graphical Lasso. The experiments of face classification task on Extended YaleB database verify that: i) graph regularization with multiclass SVM improves the performance and also leads to a sparser solution compared to L_1-norm. These approaches are the best for practical problems of blending laws (XXListe Publications 4) Next to their academic interest, our work is applied to data from competition fuel department of Total. The blending of high precision gasolines needs very precise blending laws. These laws use chromatographic data in high dimension. Yet the robustness considerations seek for simple, e.g. linear or quadratic laws. The quadratic part of these laws approximates the interactions between molecules. This expert knowledge is integrated into the model by the pattern of the initial graph between features. Sparse learning methods were implemented with a strategy for active learning as new samples enrich continuously the database. So, we are seeking for a little number of predicting features (columns of data matrix) as well as a little number of typical blends (rows of data matrix). While the sparsity of features is achieved by the minimization of a multiclass SVM with graph regularization, the sparsity of observations (a NP-hard problem) is realized by a greed search for minimizing a sub modular function of the information an observation brings. It has been shown in literature that the greed search for this class of functions guarantees bounds of performance.

P2 – Myriam Tamy. **Post Doctorat, Groupe Total. Approche multivariée de la qualité de soudure des grands ouvrages.**

Objectif. Avant chaque projet de construction impliquant des opérations de soudages, TOTAL est contrainte à une procédure de qualification de soudure et à sa validation. Cette procédure est fondée sur des tests de pré-qualification coûteux et pouvant durer plusieurs mois. L'objectif de ce travail est de construire un modèle statistique qui pourrait remplacer cette phase préliminaire. Pour cela, une base de données fondée sur l'historique de tests passés a été construite. Elle comporte des variables d'entrée et de sortie de nature hétérogène et à chaque variable quantitative est associée une « incertitude ». Des échanges avec les membres du groupe TOTAL, nous ont permis de faire des hypothèses réalistes sur les lois qui pouvaient être associées à chaque variable quantitative ainsi que leurs variances. L'objectif est de mettre au point une méthode d'apprentissage d'un modèle prédictif de la qualité de soudure en intégrant à la base de données les informations sur les lois associées à chaque variable quantitative. Le défi de ce projet est la coexistence de variables ordinales et de variables réelles. Une autre difficulté vient des variables exogènes impossibles à prendre en compte telles que les conditions météorologiques ou l'environnement (sous la mer, intérieurs de bâtiments, …). Ainsi, les modèles de prédiction doivent prendre en compte la connaissance expert, doivent être précis pour répondre aux contraintes physico-chimiques et robustes pour assurer une prédiction en présence de variables exogènes. Notre recherche s'oriente vers des méthodes d'ensembles actuellement populaires et performantes sur des problématiques de classification ou de prédiction. Elles sont basées sur l'agrégation de plusieurs classifieurs, qui sont des arbres de régression ou de classification. On considère une pondération de leurs prédictions pour prédire une valeur ou une classe d'une nouvelle instance de donnée. Concrètement, un arbre est un estimateur constant par morceaux sur des partitions disjointes de l'espace des variables d'entrées. Ces partitions sont construites par des divisions dyadiques récursives de l'ensemble des variables d'entrées qui minimisent une fonction de risque. Notre approche propose d'étendre la construction d'un arbre de régression tel que CART à des données incertaines.

P3 – Marta Bienkiewicz (en cours). The impact of affective and intentional qualities on human joint action performance.

Objectif. Post-Doctorat at EuroMov. Her current work is dedicated to the EnTimeMent project (EU Horizon 2020 FET PROACTIVE project; 2019-2022) inclusive of scientific management and running studies within the 2.3.3 Time to Sync research objective, intended to understand the impact of affective and intentional qualities on human joint action performance. Perceptuomotor group synchronization is an essential feature of human activities (clapping in an audience, walking in a crowd, music playing, sport and dance). Achieving synchronization in the group involves shared intention and perceptual interaction, but also depend on how individual motor signatures (IMS) — specific blueprints of human individuals — are assembled together to form a specific group motor signature (GMS). Theoretical hypotheses are that (i) IMS and GMS incorporate spontaneous intentional and emotional qualities – forming IEMS (Individual Emotional Motor Signature) and GEMS (Group Emotional Motor Signature), that (ii) assembling participants with different IEMS affect GEMS and group sensori-motor stability and performance, and that (iii) aforementioned qualities exist at different, and/or across, temporal scales. Mission of this project is to explore the link between multiple time scales of emotion (duration, perturbation of body systems, propagation in a group) and joint action performance (synchronization and cooperation). Thus, we assume that there is a relationship between motion and qualities such as emotion or intention, that is functional, given high redundancy of data in the motion signals, and that certain properties of movement cue towards clear

intention or one emotion type. Construction of experiments to collect data is challenging. Latin squares for experiment design are used to better cover the studied domain with a minimum number of experiences. The large dimension of data due to high recording frequencies will need sparse learning techniques to find a small number of explicative features.

M2 – Tarik Boudraa, Rémi Gaillard. **Stage de Master 2. Problèmes du Postier Chinois et Voyageur de commerce dans un graphe orienté : Application au projet WASMAN.**

Objectif. Dans le cadre du projet européen Wasman (Waste Management), une des actions pilotes consistait à organiser les tournées de ramassage porte à porte ainsi que le ramassage des points de collecte (verre, papier). Ces problèmes sont bien répertoriés dans la littérature d'optimisation dans les graphes. Il s'agit du problème du postier chinois et du problème de voyageur de commerce. Dans la pratique un certain nombre de difficultés apparaissent. La première difficulté est la création du graphe de support à partir de cartes ou autres informations gps. Une phase de simplification des données (e.g. transformation d'un rond-point en un seul sommet) était nécessaire. Une autre difficulté vient du fait que le graphe des rues est un graphe mixte orienté / non-orienté. Dans ce cas le problème bien résolu du postier chinois devient un problème combinatoire NP-difficile. La collecte des points de ramassage est le problème classique du voyageur de commerce. Ce stage M2 était consacré à la conception et la réalisation de heuristiques pour résoudre ce problème pour les communes de l'agglomération Nîmoise. L'application des algorithmes sur la zone pilote a permis de réduire de 5% en moyenne l'effort de transport. Le déploiement de ces algorithmes pour tous les circuits de ramassage permettrait d'économiser 75 000 km par an (actuellement les tournées font 1 500 000 km / an).

M2 – Adrien Gimenez. Stage de Master 2. Reconstruction des réseaux corticaux du langage dans un but thérapeutique.

Objectif. Développer une méthode innovante favorisant la capacité du cerveau à se régénérer à la suite d'une lésion. Par stimulation (interne ou externe) de certaines régions du cerveau, il est possible de reconnecter deux régions du cerveau déconnectées par la lésion. Actuellement, plusieurs modèles existent englobant les ondes circulantes, les IRMs, les EEGs, etc… L'innovation viendrait de l'utilisation d'un casque (fonctionnant par EEG) qui permettrait de stimuler des zones stratégiques du cerveau pour encourager la récupération de certaines capacités après une lésion. L'aphasie chronique (difficulté pour prononcer certains mots ou les bons mots) par exemple peut être générée par une lésion et cette méthode innovante pourrait grandement aider le sujet à exprimer ce qu'il souhaite. L'objet du stage est d'analyser les réseaux corticaux du langage chez des personnes atteintes d'aphasie chronique après un accident cérébro-vasculaire et d'identifier les zones à stimuler pour aider le cerveau à se réparer. Le travail sera concentré sur l'étude d'un seul patient dont l'entreprise dispose de données IRM. A partir de ces données IRM, les réseaux corticaux du langage devront être reconstruits. La connectivité d'un cerveau peut être représentée par un graphe. Les nœuds peuvent représenter certaines zones, parties ou régions du cerveau. Les arcs peuvent être orientés ou non ; si l'arc est orienté la région d'origine est la région qui envoie l'information ; s'il est non orienté, l'arc indique juste que les deux régions sont connectées entre elles. Les liens peuvent être regroupés dans ce que l'on appelle une matrice de connectivité dont chaque ligne représente la connexion d'une région spécifique avec toutes les autres. Grâce à cette représentation, on peut voir apparaître certaines structures à l'intérieur même du réseau général. On peut citer par exemple les communautés, qui sont des ensembles de nœuds fortement reliés entre eux. Un nœud reliant deux communautés est appelé un hub. Les techniques de théorie des graphes et analyse de données ont

permis d'identifier six régions d'intérêt qui semblent correspondre à l'activité du langage : pITG r (Inferior Temporal Gyrus, posterior division Right), pITG l (Inferior Temporal Gyrus, posterior division Left), pMTG r (Middle Temporal Gyrus, posterior division Right), MidFG l (Middle Frontal Gyrus Left), Language.IFG (L) (-51,26,2) et Hippocampus r.

M2 – Reges Oberderfer. Stage de Master 2. Automatiser et valider la mise en service du programme RAPO (Recette Ajustée Par Ordinateur).

Objectif. La conception d'une recette de mélange vérifiant des contraintes très restrictives sur les propriétés est un problème difficile qui demande plusieurs itérations de la part des opérateurs. Les produits de bases sont décrits au niveau moléculaire par leur chromatogramme caractéristique et aussi au niveau des propriétés physico-chimiques qui doivent vérifier des contraintes min / max d'un cahier de charges. Lorsqu'un produit cible est connu, il est possible de chercher une recette satisfaisante comme solution d'un problème d'optimisation. Plusieurs critères d'optimisation sont envisageables (économique, minimisation du give-away, atteindre la valeur d'une propriété, …). L'objectif de ce stage est l'interfaçage de cet optimiseur avec l'environnement du métier : les chromatogrammes des bases (communication avec une base de données en temps réel), les retours des valeurs des propriétés par le laboratoire (communication avec les analyseurs en temps réel), l'introduction des couts des bases et des contraintes sur les propriétés cible.

M1 – Rémi Nahon. **Stage de Master 1. Pas et trajectoires sur un sol connecté**.

Objectif. Le projet "HUman at home projecT" (abrégé HUT) est un projet de recherche initié en 2017 par l'université de Montpellier à la MSH (Maison Sud de l'Homme) centré sur diverses expérimentations autour d'un appartement « intelligent ». Walk@Home, un des sous-projets de HUT, propose de documenter le comportement locomoteur humain à domicile à partir notamment de la mesure des flux de marche des coHUTeurs. L'objectif de ces mesures est de permettre aux chercheurs de mieux comprendre la façon dont l'individu interagit avec son lieu de vie. Selon le parti concerné, cela correspond à différents sous-objectifs. Pour les architectes par exemple, il s'agirait de voir comment mieux agencer l'appartement, quelles sont les zones les plus occupées... Pour les sciences du mouvement, il importera de rechercher des signatures motrices, c'est-à-dire des façons de se mouvoir qui correspondent spécifiquement à un individu, pour par exemple à terme essayer de détecter des pathologies à partir de ces informations... En termes de science de données, l'objectif de ce stage est d'extraire des pas puis des trajectoires ainsi que diverses métriques telles que le nombre de pas journaliers et la proportion d'occupation moyenne de chaque pièce.

3. Projet de Recherche

« Tellement à faire ! » J'écris ces mots en écho au titre « Too much to know » du livre de Ann M. Blair (Yale University Press). Le flot d'informations auquel nous sommes confrontés dû à l'avancement des technologies est souvent accompagné par la sensation de « surcharge informationnelle ». Pourtant, cette expérience n'est pas l'apanage des temps modernes. En fait, bien avant l'ère moderne, et même dans l'antiquité, les érudits se plaignaient de la surabondance des livres et ils ont développé des techniques pour sélectionner, trier, stocker et transmettre l'information à grande échelle. Ainsi, les prémices de l'actuelle science de données remontent loin dans l'histoire de l'humanité. En suivant l'argument de Ann Blair, la quantité d'informations n'est pas la seule en cause de cette sensation de « surcharge ». Il est important de considérer le rapport entre la quantité d'information à traiter et la capacité de traitement. Aujourd'hui, la quantité d'informations « brutes » est en croissance exponentielle. Les outils mathématiques et informatiques pour traiter cette matière première connaissent un développement sans précédent.

… mais, les attentes et espoirs sur l'IA et la science de donnés sont encore plus nombreux. Ces attentes sont parfois fruit de terminologies approximatives, métaphoriques. Il suffit de lire les titres de magazines de vulgarisation telles que Science & Vie, où il ne se passe pas un numéro sans un article sur le sujet :

> Les Poubelles intelligentes ; Google Duplex invente le leurre conversationnel ; IA met le chaos K.O. ; Tableau signé d'une IA : la science se joue de l'art ; Industrie, l'innovation dopée par l'IA ; Le jour où la première IA a présenté le journal télévisé ; Séismes : l'IA redonne l'espoir de les prédire ! ; Game of Thrones : l'IA raconte la fin ; Ordinateur quantique : vers un grand mariage avec l'IA ! ; IA : l'équation qui change tout (*« S'adapter à un contexte changeant c'est le gros problème de l'IA. Une équipe vient de réussir l'exploit. Son secret ? Un système d'équations qui modélise la modulation de nos propres neurones »*) ;

Dans la réalité des applications, on rencontre des résultats encourageants mais aussi des déceptions profondes comme cela a été le cas avec les vagues précédentes de l'intelligence artificielle. Il a été reproché à John McCarthy, inventeur du terme IA, de l'avoir choisi volontairement superlatif afin de séduire investisseurs et universitaires, un choix qui serait à l'origine de déceptions ultérieures, voire de la peur. Pour Luc Julia, la discipline aurait tout intérêt à changer son nom pour 'intelligence augmentée » … ! En voici quelques titres qui vont dans ce sens (tirés de Science & Vie) :

> I.A. La faille inattendue – Quelques pixels modifiés d'une orange et l'algorithme y voit un hélicoptère ; L'IA va-t-elle bousculer l'élection américaine ? (*« Génération de texte automatique, « bots » évolutifs … les progrès de l'IA ouvrent de nouvelles possibilités au hackers malveillants pour produire des fake news toujours plus convaincants et inonder les réseaux sociaux »)* ; La voiture autonome : elle révèle qui on préférerait écraser ; Reconnaissance faciale : le grand malaise ; Le syndrome de la boite noire : l'IA porte une faille de taille, son opacité ; Données privées en voie d'extinction ; Robots soldats : la tentation d'un permis de

tuer ; Consommation : l'impasse énergétique, une débauche d'énergie silencieuse mais critique ; I.A. se prend le mur de Gödel :

> *« C'est un danger invisible mais intrinsèque aux intelligences artificielles : il est impossible de savoir avec certitude si elles feront bien ce qu'on leur a appris. Théorisée grâce aux travaux du logicien Kurt Gödel, cette indécidabilité menace l'avenir même des IA. »*

Et tout le monde y va de son mot :

- C. Villani – « L'IA n'est plus seulement un programme de recherche confiné aux laboratoires ou à une application précise. Elle va devenir une des clés du monde à venir ».
- B. Obama – « L'IA promet de créer une économie plus productive et efficace. Si elle est bien exploitée, cela peut générer énormément de prospérité et d'opportunités ».
- V. Poutine – « Celui qui deviendra leader dans ce domaine sera le maître du monde ».
- X. Jinping – « Il est nécessaire d'explorer l'utilisation de l'IA dans la collecte, la production, la distribution, la réception et le retour d'informations afin d'améliorer de manière globale la capacité à guider l'opinion publique ».
- B. Gates – « Il n'existe pas tant de technologies dans le monde qui soient à la fois aussi prometteuses et aussi dangereuses ».
- M. Zuckerberg – « L'IA améliorera notre vie à l'avenir, et le scénarii catastrophiques sont plutôt irresponsables ».
- T. Jagland – « Les conséquences de l'IA sur la démocratie restent à clarifier ».
- A. Azoulay – « L'IA est la nouvelle frontière de l'humanité. Une fois que celle-ci sera franchie, une nouvelle forme de civilisation humaine verra le jour ».
- Falque-Pierrotin – « L'objectif est de garantir que l'IA augmente l'homme, plutôt qu'elle ne le supplante ».
- E. Musk – « Je pense que l'IA est bien plus dangereuse que l'arme nucléaire ».
- J. Attali – « Il importe de maintenir la possibilité de littéralement tuer l'IA ».

Il est difficile de trouver son chemin dans cette cacophonie où se mêlent politique, économie, publicité, … ! Le plus simple est de revenir sur le chemin de transformation des données par des outils mathématiques avec, autant que possible, des preuves pour tenter de voir plus clairement la frontière entre ce qu'on peut et ce qu'on ne peut pas. Les questions et défis posés se trouvent pleinement dans le thème de recherche de mon équipe PIAS (Perception In Action & Synchronization) de l'UMR EuroMov DHM. L'objectif principal de PIAS est la découverte des lois qui gouvernent la perception humaine des agents en mouvement (Perception in Action) et la synchronisation humain-environnement en général. Cet objectif soulève des questions scientifiques interdisciplinaires et transversales : (i) extraire des régularités dans le flot de l'information qui résulte de l'interaction entre un observateur et son environnement et comprendre le rôle de ces régularités dans la réussite de la coordination du comportement : (ii) comprendre l'émergence de patterns et de la stabilité dans la synchronisation sociale en dyade et en groupe. La réponse à ces questions repose sur des approches complémentaires de santé, science du mouvement et science de données. En ce qui concerne la science de données PIAS a défini deux choix a priori sur : (i) les types d'expériences ; (ii) l'analyse des données. Pour (i), les expériences en laboratoire avec des conditions contrôlées sur un petit nombre de sujets (en bonne santé ou présentant une pathologie) seront équilibrées par un grand nombre d'expériences écologiques en conditions naturelles. Pour (ii), nous réaliserons la totalité de la chaine

de traitement de données : capture, stockage, détection de variables explicatives et redondances, réduction de dimension, statistiques, modèles prédictifs. Plus en détail, deux lignes de recherche seront suivies :

- Synchronisation dyadique : Une coalition de deux individus peut répartir et coordonner ses actions afin d'atteindre un objectif commun. Les humains se synchronisent naturellement même lorsqu'ils ne sont pas conscients de cela (synchronisation non intentionnelle). Néanmoins, il y a des cas où cette synchronisation non intentionnelle ne fonctionne pas ou le couplage est très faible. Il est important (pour aider les individus atteints de pathologies dégénératives) de comprendre les sources de ce type de synchronisation afin de proposer des remèdes pour inciter la coordination spontanée. Dans les interactions humain-machine (avatar et robot) les interactions ne sont pas parfaites et « humaines ». Un avatar qui performe les mêmes synchronisations spontanées que les humains serait mieux accepté. De façon similaire, des participants en bonne santé ne sont pas enclins à interagir avec des patients souffrants de déficits sociaux (schizophrénie, phobies sociales, autisme) parce que ces populations n'ont pas d'interaction motrice naturelle. Un de nos objectifs est d'utiliser les interactions homme-machine afin d'améliorer les interactions motrices avec des patients souffrant de déficits sociaux.

- Synchronisation de groupe : L'augmentation du nombre d'agents ($n > 2$) engendre de nouveaux défis dus au mixage des délais et l'augmentation de la dimension de l'espace de représentation. Il est possible que le groupe apporte des schémas de synchronisation qui ne sont pas visibles pour les dyades. Ces questions seront abordées à travers des manipulations expérimentales de paramètres (nombre de participants, fréquences propres, topologie (graphe) des connexions, délais, type de couplage, intention). Des points importants seront l'identification du leadership moteur, les signatures motrices individuelles ou de groupe, l'incarnation de l'émotion dans le mouvement. La synchronisation d'ensembles larges ($n = 30+$, citoyens marchant dans un espace ouvert) sera étudiée afin de mieux comprendre l'entrainement moteur social.

Une bonne partie des outils mathématiques et informatiques pour atteindre ces objectifs ne sont pas sur « l'étagère ». En plus de techniques classiques d'analyse, il sera nécessaire de maitriser et appliquer des techniques de dynamique non-linéaire, des outils de réduction de dimension (via sparse learning), des approches de graph signal processing afin d'intégrer la topologie avec les signaux temporels, de l'apprentissage actif afin d'intégrer des informations arrivant « au fil de l'eau ».

Mon projet de recherche en optimisation et science de données s'inscrit pleinement dans ces objectifs. Il est illusoire de vouloir traiter et contribuer sur tous ces aspects, le cadre est énorme par rapport aux moyens et le temps imparti à un humain ou groupe d'humains. Heureusement, le terrain n'est pas vide. Plusieurs travaux anciens, récents et en cours ont créé les prémices pour apporter des réponses aux questions posées. Les paragraphes suivants donnent un aperçu des travaux à suivre.

Optimisation Proximale et Sparse Learning

Les méthodes de descente de gradient sont un outil classique pour résoudre des problèmes d'optimisation différentiables, sans contraintes et de dimension modérée. Les algorithmes d'optimisation proximale peuvent être vus comme un outil analogue pour des problèmes non différentiables, sous contraintes, de grande dimension et distribués. Ces algorithmes sont particulièrement bien adaptés pour résoudre des problèmes d'apprentissage automatique sur des grands ensembles de données en dimension élevée. Les méthodes proximales ont un niveau

d'abstraction plus élevé que les algorithmes d'optimisation classiques (steepest descent, méthode de Newton, …). Ces derniers utilisent des opérations basiques d'algèbre linéaire, de calcul de gradient ou Hessien. L'opération de base d'un algorithme proximal est l'évaluation de l'opérateur proximal d'une fonction qui nécessite la résolution d'un petit problème d'optimisation convexe. Ces problèmes ont souvent des solutions analytiques ou peuvent être résolus très efficacement par des méthodes spécialisées. La définition de l'opérateur proximal d'une fonction convexe f: $R^d \rightarrow R$ remonte en 1962 dans un article de Moreau [1] :

$$p = prox(f, v, \beta) = argmin_x \left(f(x) + \frac{1}{(2\beta)} \|x - v\|_2^2 \right)$$

La fonction à minimiser est fortement convexe et par conséquent, le minimum p est unique. Lorsque $f(x)$ est la fonction caractéristique d'un ensemble convexe C, l'opérateur proximal fournit la projection p d'un point quelconque v sur C. Ainsi, cet opérateur peut être vu comme une projection généralisée. L'opérateur proximal est relié avec la dualité par la décomposition de Moreau, $v = prox(f, v, \beta) + prox(f^*, v, \beta)$ où f^* est le conjugué convexe de f. Cet opérateur regroupe plusieurs propriétés fondamentales de la convexité. Ce n'est pas surprenant d'obtenir en retour des résultats intéressants en termes d'existence d'optimum, de convergence et d'efficacité calculatoire (Lemaire [2]). Un excellent survey (Combettes, Pesquet [3, 4]) démontre la puissance et la flexibilité des méthodes proximales. Ils montrent qu'un nombre d'algorithmes connus (seuillage itératif, Landweber projeté, gradient projeté, décomposition alternative de Bregman, …) sont des cas spéciaux de l'algorithme proximal. De ce point de vue, le formalisme proximal fournit un cadre unificateur pour analyser et développer une grande classe d'algorithmes d'optimisation convexe. De façon analogue avec la descente de gradient, l'itération de base de l'algorithme proximal s'écrit :

$$p_{n+1} = prox(\gamma_n f, p_n, \beta) = argmin_x \left(\gamma_n f(x) + \frac{1}{(2\beta)} \|x - p_n\|_2^2 \right)$$

Un choix approprié [3, 4] des constantes γ_n dans un intervalle $\left[\varepsilon, \frac{2}{\beta} - \varepsilon\right]$ garantit la convergence de la suite ci-dessus vers un optimum.

Ces propriétés permettent des applications intéressantes. Typiquement, lorsque le nombre d'observations et bien inférieur au nombre de variables explicatives (ce qui est une caractéristique fondamentale des données de notre projet de recherche) les approches proximales permettent de choisir « presque » le meilleur sous ensemble de variables explicatives en résolvant un problème d'optimisation convexe (Candès et al. [5,6]) lequel peut être réécrit comme un problème d'optimisation linéaire. D'autres applications importantes des méthodes proximales qui résultent en solutions parcimonieuses concernent la restauration d'images (Pustelnik [6]). Récemment, dans (Pustelink et al. [7]) nous avons adapté une approche proximale afin d'apprendre les interactions entre les variables. Une amélioration de la convergence est réalisée sur la base de la projection épigraphique et une variante primal-dual de l'algorithme proximal.

[1] J.-J. Moreau, Fonctions convexes duales et points proximaux dans un espace Hilbertien, Reports of the Paris Academy of Sciences, Series A, vol. 255, pp. 2897‑2899, 1962.
[2] B. Lemaire, The proximal algorithm, International Series of Numerical, Mathematics, pp. 73‑87, 1989.

[3] P. Combettes and J.-C. Pesquet, A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery, IEEE Journal on Selected Topics in Signal Processing, vol. 1, no. 4, pp. 564‑574, 2007.
[4] P. Combettes and J.-C. Pesquet, Proximal splitting methods in signal processing, Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185‑212, 2011.
[5] E. Candès, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics 2007, Vol. 35, No. 6, 2313–2351, DOI: 10.1214/009053606000001523
[6] E. Candès, M. Soltanolkotabi, Discussion of latent variable graphical model selection via convex optimization, *Annals of Statistics*, pp. 1997‑2004, 2012.
[6] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," IEEE Transactions on Image Processing, vol. 20, no. 9, pp. 2450‑2462, 2011.
[7] M. Jiu, N. Pustelnik, S. Janaqi, Sparse hierarchical interaction learning with epigraphical projection, arXiv:1705.07817v3 [cs.LG] 18 Dec 2017.

Apprentissage Actif

La redondance des variables explicatives permet de diminuer la dimension de l'espace de représentation par diverses méthodes dont l'optimisation proximale. Un autre type de redondance est celle des observations. Il n'est pas rare de trouver dans un ensemble de données plusieurs observations qui « disent la même chose » (le fonctionnement nominal d'un process par exemple). Aussi, d'autres observations ne sont pas critiques pour la définition d'un classifieur ou modèle prédicteur. Un cas bien connu est les support vectors qui sont un sous ensemble d'observations suffisant pour déterminer la surface séparatrice des classes. Dans un cadre plus général, il s'agit de chercher les observations « informatives ». Ces observations seront cruciales dans le cas où les données arrivent « au fil de l'eau » et les modèles devront être mis à jour par des stratégies diverses d'apprentissage actif.

Plusieurs types de problèmes à résoudre se rapportent dans ce cadre. Ils ont tous une caractéristique en commun : la recherche d'une solution exacte est un problème combinatoire NP-difficile. Néanmoins, plusieurs critères estimant la quantité d'information fournie par une observation satisfont la propriété importante de la sous-modularité. Soit E un ensemble (d'observations par exemple), une fonction $f: 2^E \rightarrow R$ est sous-modulaire si :

$$\forall A, B \subseteq E, f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$$

Ou de façon équivalente :

$$\forall A \subseteq B \subseteq E, x \in E \backslash B, f(A \cup x) - f(A) \geq f(B \cup x) - f(B)$$

La sous-modularité est l'analogue combinatoire de la convexité. Intuitivement, l'ajout d'une observation x informe plus si le nombre de capteurs actuels A est plus petit que B. Plusieurs problèmes se ramènent dans ce cadre (Golovin, Krause [1]) :

- Soit le problème de déploiement d'un ensemble de capteurs afin de contrôler un phénomène spatial. On cherche le meilleur sous ensemble de k locations pour les capteurs. Dans cette application, intuitivement, l'ajout d'un capteur aide plus si le nombre de capteurs actuels est petit et aide moins si beaucoup de capteurs sont déjà déployés. Cette caractéristique de retour

décroissant est à la base de la notion de sous modularité – la superficie couverte par les capteurs est une fonction sous modulaire des placements. La variante stochastique du problème précédent est : déployer un capteur à la fois. Les capteurs peuvent tomber en panne avec une probabilité donnée. L'objectif actuel est de maximiser la superficie couverte par des capteurs valides. Ainsi, lors du déploiement du capteur suivant, nous devons prendre en compte les pannes des capteurs précédents. Ce problème a été étudié par (Asadpour, Nazerzadeh, and Saberi [2]) dans le cas où les pannes des capteurs sont indépendantes. Golovin, Krause [1] montrent que l'objectif de couverture est sous-modulaire adaptatif et ils traitent des cas plus généraux. Lié à ce problème on trouve celui du placement d'un nombre minimum de capteurs afin d'atteindre une couverture de zone maximale (Liu, Parthasarathy, Ranganathan, and Yang [3]).

- Viral Problem. Un réseau social est donné et l'objectif est d'influencer un maximum de personnes. Pour y arriver, on « contamine » un sous ensemble de personnes en espérant qu'elles contamineront leurs contacts. Kempe, Kleinberg, and Tardos [4] montrent que la fonction qui estime l'espérance des personnes contaminées est sous-modulaire. La variante stochastique est de contaminer initialement une personne, observer les personnes contaminées, et de façon adaptative choisir la personne suivante, … Une large classe de fonctions d'influence adaptative est sous-modulaire adaptative. La détection d'un sous-ensemble contaminant maximal permet de prendre des mesures efficaces d'isolement.

- Une autre application est l'apprentissage actif dans le diagnostic automatique : on a des hypothèses sur l'état du système (exemple, la maladie d'un patient) et on souhaite effectuer des tests pour identifier la bonne hypothèse. Ainsi, nous voulons choisir des exemples afin de réduire l'espace des variantes (l'ensemble d'hypothèses consistants) le plus rapidement possible. Golovin, Krause [1] montrent que la réduction dans l'espace des variantes est une fonction sous- modulaire adaptative. Cette observation est utilisée à prouver qu'un algorithme glouton adaptatif donne une politique de requête proche-optimale. Ceci généralise les résultats de (Kosaraju, Przytycka, and Borgstrom [5]).

Le rôle de la sous-modularité pour la recherche de solutions à ces problèmes est crucial. La recherche d'heuristiques efficaces est guidée par des résultats de Nemhauser [6], Minoux [7]. Ils montrent que l'algorithme glouton pour choisir le prochain $x \in E,$ qui maximise $f(A \cup x) - f(A)$ pour une fonction sous-modulaire f fournit une solution A tel que :

$$\left(1 - \frac{1}{e}\right) f\left(A_{\text{optimal}}\right) \leq f(A) \leq f\left(A_{\text{optimal}}\right)$$

Un résultat de Feige [8], montre qu'aucun algorithme de complexité polynomiale ne peut améliorer cette borne inférieure $\left(1 - \frac{1}{e}\right) \approx 0.67$.

A la lumière de ces résultats, il sera nécessaire de concentrer les efforts dans la recherche de fonctions $f(A)$ estimant la quantité d'information apportée par un sous-ensemble A. Actuellement, plusieurs fonctions mesurant l'information sont sous-modulaires : l'entropie de la distribution conditionnelle $H(A \mid B) = -\sum_{a \in A, b \in B} prob(a, b) \log\left(prob(a|b)\right.$; le gain d'information $I(B, A) = H(B) - H(B \mid A)$, les fonctions de couverture, etc.

[1] D. Golovin, A. Krause, Adaptive Submodularity: Theory and Applications in Active Learning, Journal of Artificial Intelligence Research 42 (2011) 427-486.
[2] Asadpour, A., Nazerzadeh, H., & Saberi, A. (2008). Stochastic submodular maximization. In WINE'08: Proceedings of the 4th International Workshop on Internet and Network Economics, pp.477–489, Berlin, Heidelberg. Springer-Verlag.
[3] Liu, Z., Parthasarathy, S., Ranganathan, A., & Yang, H. (2008). Near-optimal algorithms for shared filter evaluation in data stream systems. In SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 133–146, New York, NY, USA. ACM.
[4] Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137–146, New York, NY, USA. ACM.
[5] Kosaraju, S. R., Przytycka, T. M., Borgstrom, R. S. (1999). On an optimal split tree problem. In Proceedings of the 6th International Workshop on Algorithms and Data Structures, pp. 157–168, London, UK. Springer-Verlag.
[6] Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions - I. Mathematical Programming, 14(1), 265–294.
[7] Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. In Proceedings of the 8th IFIP Conference on Optimization Techniques, pp. 234–243. Springer.
[8] Feige, U. (1998). A threshold of ln n for approximating set cover. Journal of the ACM, 45(4), 634–652.

GSP – graph signal processing

Traditionnellement, le traitement du signal fonctionne sur des domaines à support continu. L'uniformité et la régularité spatiale de l'échantillonnage sont parmi les facteurs clés de sa réussite. Or, plusieurs domaines d'intérêt n'ont ni support continu ni régularité d'échantillonnage. GSP (Graph Signal Processing) [1] constitue un domaine émergeant avec des techniques nouvelles qui sont bien adaptées pour traiter des signaux en provenance de capteurs avec une topologie irrégulière. La répartition de capteurs sur une zone géographique, la configuration de relations sociales, les réseaux classiques de transport sont typiquement des graphes. Dans des domaines moins conventionnels, les stratégies d'organisation d'une équipe sportive peuvent être vues comme des graphes appris lors de séances d'entrainement ; la synchronisation d'une dyade ou d'un groupe peut se modéliser par la dynamique d'un graphe avec des topologies de connexions différentes. Le cerveau constitue une source de plusieurs graphes selon le niveau de définition des nœuds. Les neurosciences cognitives visent à élucider les mécanismes de traitement de l'information par le cerveau. Les techniques récentes de neuroimagerie permettent l'enregistrement de l'activité cérébrale tâche-orientée. Le calcul de la connectivité fonctionnelle / effective est central pour comprendre le comportement de régions du cerveau qui forment des réseaux distribués avec recouvrement partiel. Tant de domaines qui posent autant de défis scientifiques et techniques.

Classiquement, un graphe $G = (V, E)$ est considéré comme une entité statique. Plusieurs métriques permettent d'évaluer la connectivité, les intervalles (les ensembles de géodésiques-plus courts chemins), la centralité de nœuds, leur criticité pour la connexion, le flot d'informations, *small-worldness*, l'efficience globale, la modularité, etc. Il est conceptuellement facile d'ajouter la composante temporelle à un graphe statique :

$$G(t) = \big(V(t), E(t)\big), t = 1, \dots, T$$

Ainsi, l'information de chaque sommet $x(t) \in V(t)$ est une série temporelle, et la dynamique d'une arête est captée par $e(t) \in E(t)$. Lorsque la topologie $G(t)$ est constante, ce qui constitue le cas le plus simple, ce graphe dynamique peut être vu comme le produit cartésien entre le graphe statique G et l'axe temporel $I = [1, T]$:

$$G(t) = G \times I$$

Il est plus intéressant de quantifier l'évolution de la topologie (à travers toutes les métriques mentionnées) dans le temps, et de détecter ses changements qualitatifs. Une structure possible pour intégrer ceci est la concaténation (Δ) dans le temps de produits cartésiens :

$$G(t) = (G_1 \times I_1)\, \Delta \ldots \Delta\, (G_K \times I_K)$$

Cette structure est suffisamment riche pour capter les échelles temporelles de différents phénomènes longitudinaux modélisés par $G(t)$. La structure ci-dessus ne montre pas comment détecter les changements de topologie. Néanmoins, des travaux anciens et récents ont préparé le terrain. Parmi la multitude d'outils on retient le Laplacien et la Transformation de Fourier d'un Graphe.

- Le Laplacien. Le graphe $G(t)$ peut-être décrit par une matrice d'adjacence $A(t)$ telle que $A(i,j)(t)$ contient l'information de l'arête $e(i,j)(t)$. Le degré d'un sommet i est $d(i)(t) = \sum_{j \in V(t)} A(i,j)(t)$. Le Laplacien $L(t)$ de $G(t)$ est une matrice dont $L(i,i)(t) = d(i)(t)$ et $L(i,j)(t) = -1$ si $e(i,j)(t) \in E(t)$ et 0 sinon. Cette matrice est centrale dans la théorie spectrale [2] des graphes et apparait dans plusieurs contextes : analyse de la diffusion, *random walks*, le théorème de Kirchoff sur le nombre d'arbres couvrants, le leadership, la dynamique des oscillateurs couplés. Les valeurs propres et les vecteurs propres de cette matrice contiennent l'information sur la connectivité du graphe. Le premier vecteur propre d'une telle matrice est toujours $v = [1, \ldots, 1]^T$, ainsi, c'est le deuxième vecteur propre, appelé vecteur de Fiedler et la valeur propre correspondante qui fournissent la connectivité algébrique. Ce vecteur est crucial pour la partition spectrale d'un graphe. Une matrice analogue au Laplacien joue un rôle important pour le calcul de l'organisation modulaire du graphe [3]. C'est la matrice de modularité définie par $M(i,j)(t) = A(i,j)(t) - \mu(i,j)(t)$, où $\mu(i,j)(t)$ est l'espérance mathématique de la distribution de probabilité de $e(i,j)(t)$. Les valeurs $\mu(i,j)(t)$ représentent le null-model. C'est par comparaison avec ce null-model qu'on peut conclure si un sous ensemble de sommets est plus fortement inter connecté que la moyenne. Le calcul spectral peut se faire efficacement. Ceci permet de suivre dans le temps la variation de la connectivité, de la modularité ou de tout autre métrique et d'identifier les instants où des « sauts » qualitatifs se produisent.

- TFG Transformation de Fourier d'un Graphe. C'est une des pierres angulaires de GSP qui utilise le potentiel de la transformation de Fourier dans le domaine du signal de graphe $G = (V, E)$. Ici, la base c'est le signal $s : V \to R^{|V|}$. Les notions de variation lisse ou raide du signal peuvent être définies à partir de la notion bien définie de voisinage dans le graphe. Intuitivement, un signal est lisse lorsque des sommets voisins ont des valeurs de signal proches. En particulier, si $|s(u) - s(v)| \leq \alpha \times dist(G, u, v)$ avec α petit alors le signal est considéré lisse. Un indice de lisseur peut se formaliser par le Laplacien $L(G)$ (encore !) du graphe. La décomposition spectrale de $L(G) = F^T \Lambda F$ fournit une base de fonctions orthonormales F pour représenter le signal $z = F^T s$. Ces vecteurs propres jouent un rôle similaire avec les signaux exponentiels dans l'analyse de Fourier. TFG est déjà utilisée dans des études fMRI [7] et la création de décodeurs pour l'imagerie motrice. En plus de sa capacité de description et de l'existence de la TFG inverse, cette technique

permet de créer des filtres-graphiques qui sont simplement des opérateurs matriciels qui agissent sur les composantes de TFG.

[1] A. Ortega, P. Frossard, J. Kovacevic, J.M.F. Moura, and P. Vandergheynst, Graph Signal Processing: Overview, Challenges, and Applications, Proc. IEEE. 106 (5) 808‑828 (2018), DOI 10.1109/JPROC.2018.2820126.
[2] B. Bollobás, Modern Graph Theory, Springer-Verlag (1998, corr. ed. 2013), ISBN 0-387-98488-7.
[3] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, arXiv:physics/0605087v3 [physics.data-an] 23 Jul 2006.
[4] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J.-P. Onnela, Community Structure in Time-Dependent, Multiscale, and Multiplex Networks, Science. 328 (5980) 876‑878 (2010), DOI 10.1126/science.1184819.
[5] B. Ricaud, P. Borgnat, N. Tremblay, P. Gonçalves, P. Vandergheynst, Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs. Comptes Rendus Physique. Fourier and the science of today, 20 (5): 474–488.
[6] M.M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst: Geometric Deep Learning: Going beyond Euclidean data, IEEE Signal Process. Mag. 34 (4) 18‑42 (2017), DOI 10.1109/MSP.2017.2693418.
[7] W. Huang, L. Goldsberry, N.F. Wymbs, S.T. Grafton, D.S. Bassett, A. Ribeiro, Graph Frequency Analysis of Brain Signals, IEEE J. Sel. Top. Signal Process. 10 (7) 1189‑1203 (2016), DOI 10.1109/JSTSP.2016.2600859.

Applications, Dissémination, Formation

L'arrivée de nouvelles questions et sources d'information crée de nouveaux défis scientifiques et techniques mais aussi des défis d'acceptation de ces nouveautés par les scientifiques et les praticiens. Les applications sont visibles, atteignables mais les nouvelles techniques de traitement n'ont pas encore été largement éprouvées et adoptées. C'est un phénomène naturel qui suit chaque évolution ou découverte. Il faut un temps de maturation, de tests, d'échanges. Négliger ces aspects serait tuer les nouveautés dans l'œuf. Le travail dans un laboratoire de recherche avec un grand nombre d'applications médicales facilite l'application de ces nouvelles techniques. Un travail de fond pour convaincre en interne et les éditeurs et rapporteurs en externe est nécessaire. L'organisation de formations internes et individuelles est d'actualité. D'autre vecteurs importants sont les formations universitaires. Les étudiants d'aujourd'hui sont les utilisateurs de demain. Ainsi, l'introduction de ces nouveaux thèmes dans l'enseignement reste un travail permanent.

4. Liste des publications

Livre

1. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain: Semantic Similarity from Natural Language and Ontology Analysis. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2015.

Brevet

2. Stefan Janaqi, Mériam Chèbre, Guillaume Pitollat. Method and device for monitoring induced properties of a mixture of components, in particular emission properties. WIPO Patent Application WO/2015/097305.

Revues internationales avec comité de lecture

3. Alexandre Coste, Benoît G. Bardy, Stefan Janaqi, Piotr Słowiński, Krasimira Tsaneva-Atanasova, Juliette Lozano Goupil, Ludovic Marin, Decoding identity from motion: how motor similarities colour our perception of self and others, Psychological Research, February 2020, DOI: 10.1007/s00426-020-01290-8.

4. Jiu, M., Pustelnik, N., Janaqi, S. et al. Sparse Hierarchical Interaction Learning with Epigraphical Projection. J Sign Process Syst 92, 637–654 (2020). https://doi.org/10.1007/s11265-019-01478-1.

5. Vergotte, G., Perrey, S., Muthuraman, M., Janaqi, S., & Torre, K. (2018). Concurrent Changes of Brain Functional Connectivity and Motor Variability When Adapting to Task Constraints. Frontiers in Physiology, 9. https://doi.org/10.3389/fphys.2018.00909.

6. Stefan Janaqi, Mériam Chèbre, Guillaume Pitollat. Online gasoline blending with EPA Complex Model for predicting emissions, Front. Eng 2018, Vol. 5 Issue (2): 214-226. https://doi.org/10.15302/J-FEM-2017022

7. Sébastien Harispe, David Sánchez, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain, Journal of Biomedical Informatics, Volume 48, Elsevier, pp. 38–53, April 2014.

8. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi and Jacky Montmain, The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies, Bioinformatics, Volume 30, Issue 5, Oxford journals, pp.740-742, March 2014.

9. Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi, Subontology Extraction Using Hyponym and Hypernym Closure on is-a Directed Acyclic Graphs. In IEEE Transactions on Knowledge and Data Engineering, volume 24, Issue 12, IEEE computer Society Digital Library, ISSN: 1041-4347, pp. 2288-2300, December 2012.

10. Vincent Ranwez, Stefan Janaqi, Sylvie Ranwez, An $O(n \times m)$ algorithm for calculating the closure of lca-type operators. In Ars Combinatoria, Volume 104, pp. 107-128, April 2012.

11. Stefan Janaqi, Jorge Aguilera, Mériam Chèbre, Robust real-time optimization for the linear oil blending. RAIRO - Operations Research 47(4) : 465-479 (2013).

12. Stefan Janaqi, Pierre Duchet, Generator-Preserving contractions and a Min-Max result on the graphs of planar polyominoes. Ars Comb. 55 (2000).

13. Stefan Janaqi, Charles Payan, Une caractérisation des produits d'arbres et des grilles. Discrete Mathematics 163(1-3), 201-208 (1997).

14. Dezellus O., Hodaj F., Janaqi S., Chatillon C., Eustathopoulos N., Influence of evaporation-condensation in reactive spreading, Acta Materialia 50 (2002) 4727–4740, published by Elsevier Science Ltd.

Revues nationales avec comité de lecture

15. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, Mesures sémantiques basées sur la notion de projection RDF pour les systèmes de recommandation. IC 2013 : 223-238.

16. Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi, Extraction de sous-ontologies autonomes par fermeture des opérateurs hyponymie et hyperonymie. Technique et Science Informatiques 31(1) : 11-38 (2012).

Conférences internationales avec comité de lecture

17. Pascale Montreer, Stefan Janaqi, Stéphane Cariou, Mathilde Chaignaud, Isabelle Betremieux, Philippe Ricoux, Frédéric Picard, Sabine Sirol, Budagwa Assumani, Jean-Louis Fanlo: Reliability Improvement of Odour Detection Thresholds Bibliographic Data. IPMU (1) 2018, 562-573.

18. Mingyuan Jiu, Nelly Pustelnik, Mériam Chèbre, Stefan Janaqi, Philippe Ricoux, Multiclass SVM with graph path coding regularization for face classification. MLSP 2016, 1-6, 2016.

19. Stefan Janaqi, Sébastien Harispe, Sylvie Ranwez, Jacky Montmain, Robust Selection of Domain-Specific Semantic Similarity Measures from Uncertain Expertise. IPMU (3) 2014, 1-10.

20. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis. NLDB 2014, 254-257, 2013.

21. Stefan Janaqi, Jorge Aguilera, Mériam Chèbre: Prices of Robustness and Reblending in Oil Industry. MIM 2013, 180-185.

22. Stefan Janaqi, Jorge Aguilera, Mériam Chèbre, Robust real-time optimization for the linear oil blending. RAIRO - Operations Research 47(4), 465-479 (2013).

23. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain: Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems. OTM Conferences 2013: 606-615

24. Stefan Janaqi, Sébastien Harispe, Sylvie Ranwez, Jacky Montmain, Robust Selection of Domain-Specific Semantic Similarity Measures from Uncertain Expertise. IPMU (3) 2014: 1-10.

25. Sébastien Harispe, Stefan Janaqi, Sylvie Ranwez, Jacky Montmain: From Theoretical Framework to Generic Semantic Measures Library. OTM Workshops 2013: 739-742.

26. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis. NLDB 2014: 254-257.

27. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems. OTM Conferences 2013: 606-615.

28. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain, Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis. CoRR abs/1310.1285 (2013).

29. Desire Sidibé, Philippe Montesinos, Stefan Janaqi, Fast and robust image matching using contextual information and relaxation. VISAPP, International Conference on Computer Vision Theory and Applications. Barcelona, Spain, March 8-11, 2007: 68-75.

30. D. Sidibe, P. Montesinos, S. Janaqi. A Simple and Efficient Eye Detection Method in Color Images. Image and Vision Computing IVCNZ 2006, Great Barrier Island, New Zealand. November 27-29, 2006.

31. D. Sidibe, P. Montesinos and S. Janaqi. Matching Local Invariant Features: How Can Contextual Information Help? Proceedings of 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communication and Services, Maribor, Slovenia, June 2007, pp:503-506.

32. D. Sidibe, P. Montesinos and S. Janaqi. Matching Local Invariant Features with Contextual Information: An Experimental Evaluation. Electronic Letters on Computer Vision and Image Analysis, 7(1):26-39, 2008.

33. D. Sidibe, P. Montesinos and S. Janaqi. Matching Local Invariant Features with Contextual Information: An Experimental Evaluation. Electronic Letters on Computer Vision and Image Analysis, 7(1):26-39, 2008.

34. Vitaliy Feoktistov, Stefan Janaqi, Generalization of the Strategies in Differential Evolution. IPDPS 2004.

35. Vitaliy Feoktistov, Stefan Janaqi, New Energetic Selection Principle in Differential Evolution. ICEIS (Selected Papers) 2004: 151-157.

36. Vitaliy Feoktistov, Stefan Janaqi, Hybridization of Differential evolution with Least-Square Support Vector Machines, BENELEARN 2004, pp.26-31.

37. E. Alvernhe, P. Montesinos, S. Janaqi, M. Tang. Local Minimum Distance for the Dense Disparity Estimation. VISAPP, International Conference on Computer Vision Theory and Applications, February 25 - 28, 2006 Setúbal, Portugal.

38. S. Janaqi, M. Vasquez. A Tabu Algorithm for Homogeneous Partition of Samples, MIC 2001, 4th Metaheuristics International Conference, Porto, Portugal, July 16-20, 2001.

Conférences nationales avec comité de lecture

39. Stefan Janaqi, Sebastien Harispe, Jacky Montmain, Sylvie Ranwez, Sélection robuste de mesures de similarité sémantique à partir de données Incertaines d'expertise, 23ème Rencontres francophones sur la Logique Floue et ses Applications, 08-2014.

40. Quach R., Dray G., Pearson D.W., Montmain J, Janaqi S. Méthode du regroupement par soustraction étendu – LFA'2001, Rencontres Francophones sur la Logique Floue et ses Applications, Mons, Belgique, novembre 2001

Workshop

41. Sébastien Harispe, Stefan Janaqi, Sylvie Ranwez and Jacky Montmain, From Theoretical Framework to Generic Semantic Measures Library. Yan Tang Demey and Hervé Panetto eds, isbn: 978-3-642-41032-1, Springer Berlin Heidelberg, pp. 739-742, Graz, Austria, September 10-12, 2013.

Posters

42. Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi and Jacky Montmain, The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis, Lecture Notes in Computer Science, Vol. 8455, Elisabeth Métais, Mathieu Roche, Maguelonne Teisseire eds, isbn: 978-3-319-07982-0, Springer, pp. 254-257, Montpellier, France, 18-20 June 2014.

43. Parag Dharap, Sébastien Raimbault, Sylvie Arnavielhe, Gérard Dray, Stefan Janaqi, Michel Plantié, Pierre Jean, Vincent Derozier, Shubham Rastogi, Validation of Horiba Medical Pentra 80XL/XLR and Micros emiCRP Malaria Flag Performance derived from Algorithmic Data-Mining Techniques, 2015.

44. Stefan Janaqi. Quelques éléments de la théorie des graphes. Rapport de thèse, Spécialité Informatique, Option Recherche Opérationnelle, Université Joseph Fourier, septembre 1995.

Actions pour la diffusion des connaissances (hors publications scientifiques)

Stefan Janaqi. IA4DIAG - Let's medical diagnosis meet up with artificial intelligence, 9 avril 2019, Montpellier.

Stefan Janaqi. IA – IMT – L'intelligence artificielle au cœur des mutations industrielles, Colloque IMT – Mines Télécom, 4 avril 2019, Paris

5. Projets d'Application de la Recherche

Contribution Scientifique et Responsable Administratif dans des projets

Conseiller Scientifique pour le Pôle Eurobiomed

Depuis mars 2019, je contribue activement en tant que Conseiller Scientifique au Conseil Stratégique des Projets (CSP) du Pôle Eurobiomed qui a pour rôle d'assurer l'accompagnement, la labellisation et le suivi des projets innovants, ainsi que toute mission d'expertise ou d'accompagnement de projets ou d'entreprises du secteur de l'industrie et les technologies de santé. Depuis un an et demi, j'ai expertisé, conseillé, orienté plus d'une vingtaine de projets avec des budgets allant de quelques dizaines à plusieurs centaines de milliers d'euros. Quelques exemples de projets :

Acustis – AAP PIA3. Développement de solutions auditives personnalisées en ligne.

L'objectif de la société est de proposer des services d'évaluation de la perte auditive, d'adaptation des aides auditives, et de rééducation, délivrés à distance par les professionnels qualifiés. En combinant aides auditives miniaturisées et connectées, applications pour smartphone et applications professionnelles, et en se reposant sur les données d'usage et des technologies d'intelligence artificielle, le projet proposé entend améliorer l'efficacité de la correction auditive et des thérapies associées tout en abaissant significativement leur coût. Budget 250k€.

AutoBioTip – ANR. Automatisation des mesures mécanobiologiques par AFM, et de leur analyse par Apprentissage automatique.

Autobiotip a pour objectif d'éveiller l'AFM-Bio aux grands échantillons. Nous proposons une méthode d'automatisation des mesures de force pour générer des données sur au moins 1000 cellules. Le but est d'accéder à l'hétérogénéité des propriétés mécanobiologiques d'une population cellulaire. Nous recourrons, d'abord, à l'assemblage dirigé de cellules et au développement du procédé d'automatisation des mesures AFM. Puis, les analyses classiques de biophysique et les méthodes issues de l'apprentissage automatique (IA) seront adaptées pour extraire les informations mécanobiologiques issues des courbes de force (CFs). Mesurer 1000 cellules sera une avancée dans le domaine (20 cellules analysées dans les publications récentes), et l'utilisation de l'IA pour classer les CFs constitue une proposition nouvelle. Le LAAS-CNRS : AFM-Bio, Biophysics and IA, ITAVCNRS : AFM-Bio, labcom Biosoft : cell patterning et IPN-CIC : automation, rassemblent les compétences nécessaires au projet.

NovaAid (porteur Euranova) – AAP PIA3. Développement de modèles d'intelligence artificielle (IA) pour obtenir des signatures caractérisant le profil évolutif des lymphomes folliculaires (FL).

Cet outil d'aide au diagnostic pour le FL se basera sur l'ensemble des données médicales pertinentes au diagnostic (dossiers cliniques, coupes histopathologiques, et TEP) et permettra d'assister les médecins via des biomarqueurs quantitatifs (radiomics) pour caractériser le cancer et identifier la thérapie la plus appropriée.

PR1 – MEROPE – MED Interreg Project, 2002 – 2004. Douze partenaires européens. Innovative Urban Logistics Services for the Sustainability and accessibility of European cities. Interreg MED Project.

L'objectif principal de MEROPE est l'étude et le développement des modèles et instruments télématiques pour la gestion des services de logistique pour les zones urbaines et métropolitaines, avec le but de permettre le développement et l'application de nouvelles technologies de communication et information. Dans ce projet j'avais deux rôles : administratif et scientifique. J'étais chargé de la gestion du projet pour le partenaire Armines-IMT. Une des contributions scientifiques dans ce projet concerne la conception et l'implémentation d'emplacement optimal pour des services publics d'intérêt tels des parkings, blanchisseries, … Ces services génèrent une surcharge importante du trafic dans des zones urbaines saturées. Le calcul et de l'implantation optimale d'une blanchisserie dans la zone hôtelière de Rome a permis de réduire de 15% le trafic lié à cette activité. Une autre contribution concerne le contrôle d'accès aux zones restreintes par des techniques de vision par ordinateur.

PR2 – WASMAN – WASte MANagement in South Europa. MED Project, 2009 – 2012. Eight European partners.

Reinforce municipal commitment to deliver more effective Municipal Solid Waste Management (MSWM) services which are an integral part of good local governance and one of the most visible services influencing local perception of governance. I had two main tasks in this project: management and scientific. The management part concerned the financial and partnership for all French area. Our scientific contribution about the sustainability of MSWM has been assessed through eleven criteria representing various social, economic, and environmental decision-making factors. For each criterion and each area, this state-of-the-art gives the strengths and the weaknesses and a precise information about the quantity and typology of waste, its collection, segregation and processing. Based on this state-of-the-art, we assessed the MSWM processes and established a ranking of most sustainable practices within Southern Europe. At a methodological level, we provide a robust ranking that is least sensitive on the weights given to the criteria. Also, we show that few possible rankings (among a huge quantity) are realizable despite the great variety of weights given to criteria. This fact is important in its own as it demonstrates that the final ranking depends more on the objective (factual data) factors than the subjective ones (weights of criteria). (Research Report hal-00838755): A methodology for assessment of MSWM applied to Southern Europe areas). Another scientific contribution is the organization of collecting tours of waste by algorithmic methods. Good heuristics are provided to solve The Chinese Postman Problem in a digraph (NP-Hard). This point was realized by two Master 2 students at Ecole des Mines d'Alès.

Contribution Scientifique et Technique dans des projets

PR3 – RHODES 2000. (**R**épartition **HO**mogène des **D**onnées **E**n **S**ous-ensembles).

Une pratique courante en apprentissage automatique à partir des données est la partition en ensembles d'apprentissage et test. Une condition préalable à un test valide est une partition ou les deux ensembles couvrent de façon « similaire » la zone d'intérêt. Cette similarité peut être estimée avec la distance entre les densités de probabilité calculées sur chacun des ensembles. Ce calcul étant long, d'autres critères sont mis en œuvre : un critère d'inclusion, un critère d'orientation et un critère d'étendue des deux ellipsoïdes défini par chaque sous-ensemble de données via la distance de Mahalanobis. Ces trois critères ont été agrégés dans une valeur de similarité. La recherche d'une partition optimale qui maximise la similarité est un problème

combinatoire NP-difficile. Nous avons mis en œuvre deux heuristiques pour la recherche d'une bonne partition : algorithme génétique ; algorithme tabou.

PR3 – OZONE2000. Prédiction des pics d'ozone pour l'agglomération Lyonnaise.

L'objectif est de concevoir et de mettre au point des outils de prévision sur 24 heures de dépassement d'un seuil d'ozone dans l'agglomération lyonnaise. Ces outils ont été élaborés à partir des données de mesures du réseau de surveillance de la qualité de l'air géré par l'association COPARLY et de données météorologiques fournies par METEO France. Les données varient fortement d'un jour sur l'autre à cause de facteurs imprévisibles tels que les variations importantes de la circulation automobile ou d'autres phénomènes comme le transport longue distance des polluants. Trois modèles de prédiction ont été réalisés : logique floue, réseau de neurones, support vector machines. C'est ce dernier qui a été retenu pour une erreur de prédiction de 2% (sur un historique de cinq années).

PR4. – Bouygues Télécom – 2000 – 2002 – Optimisation des investissements publicitaires appliquées au marketing.

Ces investissements s'élèvent à plusieurs dizaines de million d'euros par an. Des modèles statistiques en économétrie ont établi, à partir d'analyse de données recueillies pendant plusieurs années, la forme des fonctions des retours sur investissements publicitaires unitaires. Ces fonctions ont une allure logarithmique et sont valides seulement lorsqu'un investissement a déjà eu lieu. Que fait-on lorsqu'on investit pour la première fois sur un nouveau support publicitaire ou dans un nouveau lieu ? Le choix de ces nouveaux investissements doit minimiser les risques. En même temps l'investissement sur les supports connus doit continuer tout en évitant les effets de saturation (zones de faible croissance des fonctions logarithmiques). Ce problème se traduit en optimisation à variables mixtes et plusieurs scénarii ont été établis et calculés.

PR5 – Varel – 2002 – 2003 – Méthodes de fouille de données pour l'optimisation des forages pétroliers.

Les forages pétroliers fournissent une grande quantité de macro données (logs). L'analyse de ces données est un problème bien connu par l'exploration des gisements. Ces analyses permettent de connaitre le potentiel de production d'un site. Les prédictions sont souvent grossières et des erreurs d'appréciation engendrent des pertes importantes. Des nouvelles techniques d'analyses permettent d'estimer finement la saturation en huiles de petites lamelles (micro données). Il est alors nécessaire de combiner ces deux sources d'information pour une meilleure prédiction du site. Ce problème pose de vrais défis par la taille de données et surtout par les incertitudes très variables entre les données de logs et les données de saturation de lamelles venant des laboratoires. Les modèles mathématiques qui en découlent sont des problèmes inverses souvent mal posés et leurs solutions numériques sont instables. La modélisation comme problème d'optimisation robuste a permis d'obtenir des résultats encourageants.

PR6 – PMETL – 2002 – 2003 – Etat de l'art sur les différentes méthodes de fouille de données et d'aide à la décision dans le domaine du transport. En collaboration avec le Ministère des Transports.

Dans un département fortement rural et avec une fréquentation des transports en commun à caractère saisonnier, il est nécessaire d'organiser la prise à domicile et le rabattement des usagers sur des points de concentration, départs ou arrêts d'une ligne régulière de transport en commun. Tout le trafic de rabattement doit être organisé autour de la grille horaire des

transports en commun et doit s'adapter à sa modularité saisonnière ou conjoncturelle. Il est nécessaire de dimensionner la capacité des véhicules de rabattement. Il faut étudier les conséquences du couplage entre le système de rabattement et la ligne régulière afin de réduire les points d'arrêt compressant d'autant la durée de chaque liaison. L'objectif est de rendre le réseau de transport public plus compétitif au regard de son principal concurrent : la voiture particulière. Ceci améliorerait l'utilisation des subventions, elles ne serviraient plus à faire circuler des véhicules vides mais à amener des passagers vers la ligne régulière qui nécessite moins de soutien.

PR7 – VISAMEL – 2004 – 2006. Société Total. Aide de la conduite de mélanges en ligne.

VISualisation de MELanges pour la production de produits de l'industrie pétrolière. Les mélanges actuels utilisent autour de vingt produits de base et respectent jusqu'à quarante propriétés technologiques et normes réglementaires. Ces normes deviennent de plus en plus restrictives et le calculateur de mélange en ligne est appelé toutes les 4 secondes afin d'ajuster le mélange. Dans ces conditions, il devient important pour les opérateurs de pilotage de suivre visuellement les changements de la zone faisable, les mélanges nominaux, les mélanges optimaux (selon plusieurs critères), la saturation de contraintes, etc. Une nouvelle écriture du régulateur en ligne et de la géométrie sur les polytopes de mélanges ont permis de visualiser avec précision l'évolution du mélangeur en ligne dans l'espace des propriétés ainsi que l'espace des recettes. Pour le cas de mélanges infaisables, un modèle mathématique de modification minimale des matrices de contraintes est conçu et implémenté. Ce module permet d'effectuer la meilleure correction de bornes ou de changement de base pour atteindre des solutions faisables.

PR8 – VISAMUL – 2008 – 2009. Société Total. Aide de la conduite de procédés en ligne.

La conduite de procédés industriels dépend souvent de la connaissance d'une matrice de gain qui permet de décrire localement l'évolution du procédé par des équations linéaires. Les variables de décision varient dans un parallélépipède et les matrices de gain (jacobien de la fonction de procédé) sont connues pour varier dans des intervalles min/max. Il faut alors déterminer une matrice de gain optimale et aussi visualiser la zone de fonctionnement et les valeurs nominales des propriétés. Le modèle mathématique utilise un développement de premier ordre. Ainsi, la zone de fonctionnement est localement un zonotope (image d'un parallélépipède par une transformation linéaire). Cette classe de polytopes a des propriétés remarquables (e.g. un centre de symétrie calculée analytiquement) qui permettent d'optimiser la matrice de gain et de visualiser le procédé. Une forme de modèle inverse est modélisée qui permet de calculer pour chaque variable de décision la marge de manœuvre associée. Ceci est d'une grande utilité dans la conduite des procédés.

PR9 – VADEQUA – 2012 – Prédiction de facettes de satisfaction, aide à la décision pour les ressources humaines.

L'objectif de ce projet était d'étudier et développer un outil de prédiction de satisfaction d'un candidat en fonction de son profil psychologique et des réponses aux questionnaires entreprise. Il s'agit de détecter des facettes psychologiques et des facteurs sociaux prépondérants qui déterminent le niveau de satisfaction d'un candidat à un poste donné. Cet outil sera utilisé pour aider les responsables de ressources humaines au choix et affectation des candidats. Il s'agit de prédire les scores de satisfaction vus sous deux angles différents : la personne et le recruteur. Les variables explicatives sont calculées à partir de réponses aux questionnaires spécialisés. Ces

variables sont par nature incertaines ce qui oriente le choix des modèles de prédiction vers les modèles robustes.

PR10 – ACRETION – 2014 – Serious Game – Aide à la décision aux ressources humaines.

Acrétion est un jeu vidéo multijoueur en application web. Il est composé de deux espaces de jeu : le système planétaire et la nébuleuse. Le système planétaire est propre au joueur et privé. La nébuleuse est un espace multijoueur avec interactions en temps réel. Chaque joueur possède un compte sur la plateforme Acrétion regroupant les informations qui lui sont personnelles et permettent l'interaction avec les autres joueurs ainsi que la persistance de ses données. Le joueur devra maitriser la visibilité de ses données. Acrétion regroupe des caractéristiques propres aux jeux vidéo multijoueur en ligne et aux réseaux sociaux. Les modèles mathématiques pour répondre à ces problématiques calculent l'influence et sa diffusion dans un graphe, les phénomènes viraux, le calcul des hubs (influenceurs), le calcul des barrières définissant des clusters.

PR11 – CONCOURS GE – 2014 – Analyse de données et aide à la décision pour la classification des grandes écoles.

Dans le but d'estimer le niveau réel de la sélectivité et de la notoriété des écoles d'ingénieurs, il est nécessaire de déterminer l'état de la concurrence, les facteurs d'influence de choix des candidats. Quelle est la nature du lien entre le choix du candidat et les palmarès des écoles qui paraissent chaque année dans la presse spécialisée ? Est-ce que l'importance de ces facteurs varie selon le type de population ? L'analyse des données des listes de vœux des candidats a permis d'établir une « vérité » terrain qui intègre des facteurs objectifs (les résultats scolaires) ainsi que des facteurs d'influence (la famille, la presse). A partir de ces analyses un graphe de voisinage des écoles, telles qu'elles sont vues par les candidats est établi. Ce voisinage est corrélé avec le ranking des écoles par la presse ou les organismes internationaux.

PR12 – EPAOnLine – 2014 – 2015. Calcul de mélanges avec les contraintes environnementales en temps réel. En collaboration avec la société TOTAL.

Sur la base des rejets toxiques du parc automobile, un modèle empirique Complex Model a été développé par U.S. Environemental Protection Agency. Ce modèle calcule les rejets toxiques d'une essence à partir de ses propriétés chimico-physiques. Ces propriétés vérifient des contraintes et sont le résultat de mélange de bases. L'implémentation de ce modèle dans les programmes de mélanges est difficile et les implémentations actuelles utilisent des modèles de programmation à variables mixtes. Le temps d'exécution (quelques minutes) de ces modèles à base de contraintes disjonctives est prohibitif pour l'optimiseur de mélanges en ligne (un appel toutes les quelques secondes). Une nouvelle re-écriture du Complex Model sans variables binaires ou entières permet d'introduire une fonction objectif différentiable. L'optimisation de cette fonction objectif avec les autres fonctions qui contrôlent les propriétés des mélanges permet de contrôler les émissions en temps réel. Cette technique a été brevetée en 2015.

PR13 – HORIBA – 2015 – Etude de la définition d'un signal «malaria» par fouille de données des variables produites par trois outils diagnostics d'Horiba Medical.

Horiba Medical dispose de trois types de machines diagnostic capables d'analyser des échantillons sanguins humains afin de détecter la présence du parasite causant la malaria. Lors de ces diagnostics un grand nombre de paramètres ($\sim 1.2 \times 10^5$) sont évalués sur un

échantillon de $\sim 5 \times 10^2$ patients. L'objectif du projet est de trouver un petit ensemble de variables explicatives parmi les paramètres évalués. Un cas typique d'application médicale avec peu d'échantillon en dimension élevée. L'application successive de sparse learning a permis de prédire avec une erreur acceptable la pathologie avec un faible nombre de variables explicatives (quelques centaines). Ces variables sont choisies parmi les variables d'origine. Par conséquent, elles sont interprétables par les experts. Ces modèles de prédiction sont intégrés dans les machines diagnostic actuellement en fonctionnement H24 dans des zones où la compétence médicale et biologique est absente.

PR14 – NODEA – 2016 – Analyse de données pour la classification de prélèvement de biopsie.

L'objectif est de classer les cancers à partir de données en dimension élevée ($\sim 2 \times 10^3$). Il s'agit d'apprendre un classifieur à partir de données et surtout d'extraire des variables explicatives compréhensibles par l'expert. Un modèle sparse learning a permis une bonne prédiction avec une réduction significative du nombre de variables explicatives.

PR15 – SOUDQuality – 2016 – 2017. Prédiction de la qualité de soudure des grands ouvrages par machine learning. En collaboration avec la société TOTAL.

Un certain nombre de relations a été établi expérimentalement entre les paramètres de soudure et la qualité du résultat. La certification des soudures pour les grands ouvrages est un travail d'expertise long et couteux. L'objectif de ce projet est de prédire la qualité de soudure à partir de modèles basés sur l'apprentissage machine. Le défi de ce projet est la coexistence de variables ordinales et de variables réelles. Une autre difficulté vient des variables exogènes impossibles à prendre en compte telles que les conditions météorologiques ou l'environnement (sous la mer, intérieurs de bâtiments, …). Ainsi, les modèles de prédiction doivent prendre en compte la connaissance expert, doivent être précis pour répondre aux contraintes physico-chimiques et robustes pour assurer une prédiction en présence de variables exogènes.

PR16 – ENTIMEMENT – 2018 – 2022. FETPROACT European Project. Ten European Partners. Qualitative ENtrainment and synchronization in multiple TIMEs MENTal and computational paradigms. Data Sciences in the service of Human Mouvement.

Time to Sync research objective intend to understand the impact of affective and intentional qualities on human joint action performance. Perceptuomotor group synchronization is an essential feature of human activities (clapping in an audience, walking in a crowd, music playing, sport and dance). Achieving synchronization in the group involves shared intention and perceptual interaction, but also depend on how individual motor signatures (IMS) — specific blueprints of human individuals — are assembled together to form a specific group motor signature (GMS). Theoretical hypotheses are that (i) IMS and GMS incorporate spontaneous intentional and emotional qualities – forming IEMS (Individual Emotional Motor Signature) and GEMS (Group Emotional Motor Signature), that (ii) assembling participants with different IEMS affect GEMS and group sensori-motor stability and performance, and that (iii) aforementioned qualities exist at different, and/or across, temporal scales. Mission of this project is to explore the link between multiple time scales of emotion (duration, perturbation of body systems, propagation in a group) and joint action performance (synchronization and cooperation).

PR17 – MEDTRUC – 2019 – Optimisation de tournées et planification de camion médical pour servir les déserts médicaux.

Le besoin d'apporter le soin dans les déserts médicaux est croissant. L'évolution démographique implique l'évolution et l'adaptation des structures vitales de la société. La

proposition Medtrucks s'inscrit dans cette dynamique et sera intégrée en deux temps : (i) apporter le soin médical vers les personnes à mobilité réduite (personnes âgées, situation de handicap, …) ; (ii) apporter le soin médical dans les zones sous dotées selon la définition des ARS. Medtrucks propose l'utilisation de soins itinérants à travers UMM – Unité Médicale Mobile, afin de répondre aux besoins (i) et (ii) ci-dessus. Ici apparaît le besoin d'optimiser les itinéraires afin d'augmenter le ratio « temps soins / temps trajets » dans ce nouveau système. Le modèle mathématique pour répondre à la problématique contient un mélange d'un problème d'ordonnancement avec le voyageur de commerce.

PR18 – STELLA Surgical – 2020. Prédiction de la stéatose à partir d'images.

Les méthodes d'estimation de la stéatose étant loin d'être « idéales » pour de nombreuses raisons (invasive, utilisant des radiations, subjectives, peu applicables, etc), il est indispensable de développer une nouvelle méthode fiable, rapide, portable, simple et non invasive permettant d'évaluer la présence et la sévérité des lésions de stéatopathie. Nous proposons d'évaluer la performance de ces systèmes de classification selon les valeurs de stéatose, à partir d'images acquises à l'aide d'un simple smartphone mais selon un protocole d'acquisition précis. Ces images seront associées aux valeurs du bilan biologique du donneur cadavérique. La méthode de référence utilisée pour l'évaluation de la stéatose sera l'analyse histologique d'une biopsie hépatique à la fin de la transplantation hépatique. L'objectif principal de ce projet est donc d'identifier et de tester des méthodes d'analyse d'images, d'apprentissage automatique et d'intelligence artificielle capables de classifier selon les principaux pourcentages de stéatose les donneurs à travers l'utilisation des images intra opératoires du greffon hépatique et des données biologiques du donneur.

PR19 – BRAMS – 2018. Marqueurs de Parkinson dans les données BAASTA.

Le BRAMS, Laboratoire international de recherche sur le Cerveau, la Musique et le Son, est un centre unique, situé à Montréal et conjointement affilié à l'Université de Montréal et McGill University. Le laboratoire axe principalement ses recherches sur la cognition musicale, avec une emphase particulière sur les neurosciences. Le BRAMS, qui compte aujourd'hui plus de 35 membres de renommée internationale et une centaine d'étudiants et chercheurs postdoctoraux, est depuis avril 2011, part intégrante du nouveau Centre de Recherche sur le Cerveau, le Langage et la musique – CRBLM. La collaboration sera centrée sur les patients souffrant de la maladie de Parkinson (MP) et présentant des déficits dans la sphère du rythme. Ils ont des difficultés dans la discrimination perceptuelle de rythmes auditifs et dans la synchronisation du mouvement avec des stimuli rythmiques (métronomes ou musique). Cependant, nous ne savons pas jusqu'ici si ces déficits peuvent contribuer à distinguer différents profils de troubles du mouvement dans la MP, tels que la présence ou l'absence d'enrayage cinétique (freezing). Un protocole a été établi pour passer des batteries de tests (BAASTA) à plusieurs groupes de patients avec ou sans enrayage cinétique. Les données recueillies s'inscrivent dans le cadre de big data par leur quantité, typologie et source d'acquisition. L'extraction des informations utiles et pertinentes pour détecter les différents profils de troubles à partir de ces données est une tâche classique et pourtant difficile en science de données. La conception et la mise en œuvre de méthodes mathématiques et statistiques d'analyse adaptées au problème en question restent à faire.

6. Enseignement

Modules Enseignés en Ecoles d'Ingénieurs et Universités

Mathématiques pour l'Apprentissage Machine

Niveau M2

OBJECTIF : Présenter le lien entre le Machine Learning et l'Optimisation Convexe.

Clarifier l'apport des techniques d'optimisation sur un certain nombre de problématiques liées à l'apprentissage automatique. Les support vector machines sont basées sur un sous ensemble critique d'observations. Ce sous ensemble correspond dans l'espace dual aux variables duales égales à 0. La réduction de la dimension sans passer par des projections dans des espaces abstraits (ACP par exemple) nécessite l'optimisation avec pénalisation de la norme L_1.

Moyens pédagogiques : Cours, TD, TP et projet (50h)

Support pédagogique : polycopie, librairies de codes.

Challenges Data Science IMT.

Niveau M1

OBJECTIF : Résoudre des problèmes réels de Data Science en compétition avec toutes les écoles du groupe IMT.

Challenge 1 : Prédiction de la performance d'un puits.

Challenge 2 : Implantation optimale de stations de service.

Moyens pédagogiques : Librairies de code et réalisation de projet (50h)

Statistiques et probabilités avancées

Niveau M1

OBJECTIF : Renforcer les compétences théoriques avec un approfondissement des statistiques et de la théorie des probabilités.

Le cadre PAC (Probably Approximately Correct) pour les algorithmes de machine learning. Les « pièges » qui guettent les modèles : univers de données inconnu, malédiction de la dimension, conditions suffisantes Lipschitziennes pour la convergence des algorithmes, compromis biais-complexité. Techniques de réduction de la dimension.

Moyens pédagogiques : Cours, TD, (50h)

Support pédagogique : polycopie.

Recherche Opérationnelle et Aide à la Décision I

Niveau M1

OBJECTIF : Initiation aux méthodes classiques de recherche opérationnelle et de la décision multicritère.

Introduction à la programmation linéaire, modélisation d'un problème comme problème d'optimisation, résolution par la méthode simplexe. Introduction à la théorie des graphes et quelques algorithmes de base (plus court chemin, arbre min). Introduction à la complexité combinatoire. Initiation à la décision multicritère.

Moyens pédagogiques : Cours, TD, TP et projet (70h)
Support pédagogique : polycopie, librairies de codes.

Recherche Opérationnelle et Aide à la Décision II

Niveau M2
OBJECTIF : Méthodes avancées de recherche opérationnelle et de la décision multicritère.
Introduction de la dualité en programmation linéaire et son interprétation. Les flots dans les graphes. Flot max et coupe min, exemple classique de problèmes duaux, résolution efficaces par algorithme d'Edmonds, couplages dans les graphes. Problème du postier chinois. Introduction aux méthodes de surclassement en décision multicritères.
Moyens pédagogiques : Cours, TD, TP et projet (70h)
Support pédagogique : polycopie, librairies de codes.

Théorie des graphes et complexité

Niveau M1
OBJECTIF : Fondements de la théorie des graphes et introduction à la complexité des problèmes combinatoires.
Invariants des algorithmes de base en théorie des graphes. Recherche d'information globale à partir d'informations locales. Implémentations efficaces des algorithmes de base (plus court chemin, arbre min, flot, couplage, postier chinois). Heuristiques pour problèmes d'optimisation combinatoires NP-Difficiles.
Moyens pédagogiques : Cours, TD, TP et projet (50h)
Support pédagogique : polycopie, librairies de codes.

Analyse Numérique

Niveau L3 (première année d'école d'ingénieur)
OBJECTIF : Fondamentaux des algorithmes numériques.
Recherche numérique de racines, résolution itérative de systèmes d'équations linéaires, résolution d'équations à dérivées partielles avec des méthodes forward et backward, amélioration de convergence des méthodes numériques (méthode de Crank-Nicholson). Développement en Matlab.
Moyens pédagogiques : Cours, TD, TP et projet (50h)
Support pédagogique : polycopie, librairies de codes.

Calcul Scientifique

Niveau L3 (première année d'école d'ingénieur)
OBJECTIF. Introduction au calcul scientifique. Représentation de nombres, précision et overflow, problèmes numériques indésirables. Matrices et opérations matricielles, interpolation, polynômes, optimisation, équations différentielles.
Moyens pédagogiques : Cours, TD, TP et projet (50h)
Support pédagogique : polycopie, librairies de codes.

Optimisation Continue

Niveau L3 (première année d'école d'ingénieur)
OBJECTIF : Fondamentaux des méthodes d'optimisation.
Optimisation de fonctions différentiables avec ou sans contraintes. Conditions d'optimalité, dualité, conditions de Karush-Kuhn-Tucker. Méthodes de descente de gradient, steepest-descent, Newton, sub-gradient.

Moyens pédagogiques : Cours, TD, TP et projet (30h)
Support pédagogique : polycopie, librairies de codes.

Ordonnancement

Niveau M1.
OBJECTIF : Fondamentaux des méthodes d'ordonnancement.
Différentes classes de problèmes d'ordonnancement dont la plupart sont des problèmes d'optimisation combinatoires NP-difficiles. Méthodes exactes et heuristiques pour la résolution. Notions de complexité combinatoire. Ordonnancement sur un processeur, job-shop, avec ou sans préemption.
Moyens pédagogiques : Cours, TD, TP et projet (50h)
Support pédagogique : polycopie, librairies de codes.

Architecture d'ordinateur et logique mathématique.

Niveau L3 (première année d'école d'ingénieur.
OBJECTIF : Structure de base d'un processeur.
Représentation de nombres, bus, registres, fonctionnement de base d'un processeur. Logique mathématique de premier ordre, réduction conjonctive et disjonctive d'expression logiques.
Moyens pédagogiques : Cours, TD (50h).

Recherche Opérationnelle et Aide à la Décision – Niveau 1

CNAM, Module RCP 101, Niveau L3 (première année d'école d'ingénieur)
OBJECTIF : Techniques de base d'optimisation linéaire et théories graphes.
Moyens pédagogiques : Cours, TD, TP et projet (50h) - Support pédagogique : polycopie, librairies de codes

Recherche Opérationnelle et Aide à la Décision – Niveau 2

CNAM, Module RCP 110, Niveau M1 (deuxième année d'école d'ingénieur)
OBJECTIF : Programmation et théorie de graphes avancée.
Moyens pédagogiques : Cours, TD, TP et projet (50h).
Support pédagogique : polycopie, librairies de codes.