FICHE PROJET EUROPEEN			
ACRONYME : DAISI			
NOM COMPLET DU PROJET	DAISI Data Data science : méthodologies et applications		
NUMERO DE CONVENTION	HN0005604		
DATE DE DEBUT	01/10/2016		
DATE DE FIN	30/03/2020		
COORDINATEURS	Laurent HEUTTE		
• *Etablissement(s)*	• *Laboratoire(s)*	• *Responsable(s)*	• *Partenaire(s)*
	LITIS		
CONTACT			
SITE INTERNET DU LABORATOIRE ET PROJET			
DESCRIPTION DU PROJET			
RESUME	*Contexte, présentation générale de l'opération :* L'ubiquité des espaces de stockage de données numériques et la facilité d'acquérir et de générer de nouvelles données, via des smartphones, des tablettes, des caméras embarqués, des caméras fixes de surveillance, des systèmes de suivi de type GPS, une carte Vitale, des systèmes d'imagerie médicale etc., permettent à tout un chacun et à des organismes publics de stocker une quantité phénoménale de données, comme en témoignent les bibliothèques numériques chargées de photos et de vidéos personnelles, les informations stockées par les réseaux sociaux, ou les portails open data récemment mis à disposition par les régions et l'état français. Les données typiquement générées et stockées peuvent être de natures hétérogènes et variées : par exemple, elles peuvent être relatives à des images de poumons d'un patient atteint d'une maladie, aux habitudes d'achats des clients d'un magasin ou aux renouvellements des marquages sur les routes du département de l'Eure. L'ensemble de ces données forme une mine de richesses qu'il est nécessaire d'explorer, de nettoyer et d'interpréter afin de générer de la connaissance. Comme le souligne également le rapport Mc Kinsey 2011, cette capacité à extraire de la connaissance à partir de données sera à très court terme, et pendant longtemps, un vecteur d'innovations et de richesses. De ce fait, les découvertes scientifiques obtenues par analyse de données sont de plus en plus nombreuses et interviennent dans plusieurs champs disciplinaires tels que la biologie, la santé, les sciences humaines et sociales. … Dans ce contexte, la science des données (data science) est le moteur central de ces recherches et de ces innovations et les questions scientifiques qui y sont abordées traitent des problèmes d'acquisition, de stockage, d'indexation, de modélisation et d'analyse. Parmi les questions		

	difficiles, on s'intéresse typiquement à l'intégration de ces données complexes, hétérogènes et interdépendantes pour induire de la connaissance, aider à la prise de décision ou générer de la valeur. La région Haute-Normandie et l'ensemble des acteurs institutionnels en région, pleinement consciente de l'importance de la filière numérique, multiplient ses actions de développement et ses soutiens envers les acteurs et entrepreneurs locaux, générant ou traitant des masses de données numériques ; citons par exemple, la création de Seine Innopolis et de la cantine numérique. Afin de faire levier sur ces données, il est donc de première importance d'amorcer une structuration des activités locales liées à l'analyse et la science des données. Le projet que nous proposons ici est une opportunité unique de lancer et de fédérer les activités autour de ce thème. Le projet DAISI vise à structurer une partie des travaux menés en sciences des données et impliquant différents acteurs haut-normands (voir la liste des partenaires ci-dessous). Son objectif est de fédérer les forces et les moyens acquis et mis en œuvre par ces acteurs afin de renforcer les compétences existantes, de faire émerger des nouveaux champs de recherches nés de la fertilisation croisée de domaines scientifiques particuliers et de la science des données, pour finalement amener les acteurs à produire des résultats scientifiques de portée internationale de par leurs aspects innovants. D'un point de vue scientifique, son objectif est de développer des méthodes et des outils innovants afin d'analyser des grandes quantités de données, d'en extraire des informations et connaissances utiles pour différents champs disciplinaires telles que la santé, l'environnement urbain et l'analyse de trafic urbain. De ce point de vue, DAISI est donc un projet de recherche amont guidé par des applications à fort impact sociétal.
OBJECTIFS	*Objectifs recherchés, résultats escomptés et public visé :* Les objectifs des travaux de recherche auxquels s'attaque DAISI relèvent de plusieurs problématiques : • Dans la plupart des cas d'usage qui nous intéressent, les données nous parviennent sous une forme brute, brute dans le sens où elles ne sont pas structurées, elles sont complexes et hétérogènes et leur lecture directe n'est que très faiblement informative. Un de nos premiers objectifs sera donc de transformer ces informations brutes sous une forme permettant une meilleure interprétation quantitative et qualitative. Afin de faire émerger cette nouvelle forme de représentation, nous viserons à développer des méthodes exploitant les structures (inconnues et sous-jacentes) des données à disposition. Les cadres applicatifs de ces nouvelles méthodologies développées au sein de DAISI seront liés à des problématiques de santé via l'analyse de données hétérogènes (images, génomiques et cliniques) et à des problématiques d'analyse d'environnement urbain à travers un signal sonore.

- L'omniprésence et la miniaturisation des systèmes d'acquisition a permis de simplifier à l'extrême la capture d'information. Ainsi, il est de plus en plus commun d'obtenir des informations concernant un système, non plus statiques, mais évoluant au cours du temps. Si une majorité des méthodes actuelles liées à la science des données se focalisent sur des informations statiques, un verrou scientifique important est posé par l'analyse des données séquentielles. La difficulté du problème scientifique sous-jacent est d'autant plus grande lorsque, comme dans certains de nos cas d'usage, les données se présentent sous une forme fortement structurées, telles que des séquences de graphes ou des séquences de marqueurs dans un plan 2D ou de manière générale des séquences d'objets complexes. Dans ce cadre d'une problématique de fouille de séquences complexes, DAISI visera à développer des méthodes génériques permettant de faire jaillir de la masse de séquences, des séquences caractéristiques, des séquences anormales ainsi que des motifs (ou des sous-séquences) distinctifs de ces séquences. Au sein de DAISI, ces problématiques font référence à des cadres applicatifs liés à l'analyse de la mobilité dans un cadre de trafic routier et de personnes et également à l'analyse de mouvements et de la posture avec pour objectif final la définition d'une rééducation personnalisée pour les personnes atteint de pathologie de la régularisation posturale.

- Dans un contexte d'analyse prédictive, une des tâches les plus fréquentes est celle d'attribuer une catégorie donnée aux observations associées à une donnée. Dans des cas simples, cette tâche appelée discrimination ou catégorisation peut être réalisée grâce à des outils technologiques issus de la recherche scientifique. Cependant, si les données à disposition sont représentées à partir d'observations complexes, hétérogènes mais interdépendantes, la question du choix de la meilleure observation ou de la meilleure façon de combiner ces différentes "vues" d'une même donnée peut se poser. Dans le cadre de DAISI, nous nous attaquerons à plusieurs questions relatives à la discrimination et notamment à la discrimination à partir de plusieurs vues. En particulier, nous viserons à développer de nouveaux cadres méthodologiques, basés sur les méthodes parcimonieuses, permettant de prendre en compte la diversité et les complémentarités des différentes vues. Dans le cadre du projet, ces problématiques émergent notamment dans les applications médicales pour la cancérologie où les données relatives à un patient peuvent provenir de différentes vues (imagerie médicale, données cliniques, génomiques, …). Nous investiguerons également l'apport de méthodes de type forêts aléatoires, à base d'arbres de décision, qui sont particulièrement adaptées au traitement de ces données hétérogènes de par leur provenance mais aussi leurs représentations.

Outre des développements méthodologiques, dont la validation se fera par le biais de publications scientifiques de haut niveau et par le

	développement de paquets logiciels, DAISI propose des cadres applicatifs qui nécessitent une adaptation et une mise en oeuvre appropriée des méthodes développées. L'objectif des problématiques liées à la science des données est l'émergence de connaissances et de savoirs, puis de valeurs dans un champ applicatif donné. Au sein de DAISI, on s'intéressera à : • l'analyse du trafic routier par fouille de séquences de graphes de l'environnement d'un véhicule • l'analyse des déplacements urbains de piétons par fouilles de signaux audio • l'analyse prédictive pour la cancérologie par association de la radiomique et de la génomique • l'analyse de régulation de la posture
IMPACTS ATTENDUS ET FINALITE	*Impacts attendus-diffusion et capitalisation des résultats :* • Rapport technique : Méthodologies d'analyse de séquences avec application à l'identification des différents modes de régulation posturale au niveau intra- et inter-individuel • Rapport technique : Méthodologies d'apprentissage de représentation à l'aide de réseaux de neurones profonds et d'apprentissage de dictionnaires • Rapport technique : sélection de caractéristiques, l'apprentissage supervisé et la fouille dans les espaces de dissimilarités pour la catégorisation de données hétérogènes • Rapport technique : Mise en place d'un protocole d'évaluation de la régulation posturale en situation normale, contrainte (en termes de fréquence d'oscillation, mobilité articulaire et surface d'appui) et perturbée (modification brutale des contraintes avant retour à la normale). • Rapport technique sur la représentation de données cliniques, génomiques et images dans les espaces de dissimilarités • Rapport technique sur l'apport des techniques d'apprentissage profond pour le SLAMMOT dans le cadre de scènes routières. • Rapport technique sur la reconnaissance d'évènements sonores à l'aide de réseaux de neurones profonds. • Rapport technique : Méthodologie de détection de changement, évaluation du temps de relaxation dans l'analyse de la posture • Brique logicielle de SLAMMOT intégrant les modifications apportées par l'apprentissage profond. • Démonstrateur d'analyse, de fouille et de classification de données images, cliniques • Démonstrateur de logiciel ou systèmes embarqués de reconnaissance de signaux sonores urbains
RESULTATS	
MODALITES DE FINANCEMENT	
BUDGET TOTAL	2 042 365€
• *Niveau de soutien FEDER / FSE / FAEDER*	980 335 €
• *Niveau de soutien région*	90 000 €
• *Niveau de soutien Etat*	183 000 €

• *Autofinancement*	
• *Autre*	90 000 €
NOMBRE D'ALLOCATIONS DOCTORANTS	1
NOMBRE D'ALLOCATIONS ET POST-DOCTORANTS	0
L'Europe s'engage en Normandie avec le Fonds Européen de Développement Régional	