Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

Dossier de création de master

Partie A : Présentation générale de la formation

NOM DU MASTER : Mathématiques pour la science des masses de données
(Acronyme: Data Science)

Etablissement déposant de la mention
Nom : : Ecole Polytechnique
Adresse : Route de Saclay, 91128 Palaiseau

I. FICHE D'IDENTITE DE LA FORMATION

- Finalité (R ou P, R et P) : R et P
- Modalités d'enseignement (formation initiale ou continue, en alternance, par apprentissage ou contrat de professionnalisation)
  - Formation initiale
- Responsable de la formation
  MOULINES, Eric, Professeur, Télécom ParisTech
- Etablissement « Déposant » (responsable du dépôt du dossier)
  - Ecole Polytechnique
- Etablissement (s) cohabilité(s)
  - Ecole Polytechnique
  - Télécom ParisTech
- Etablissement (s) partenaires (s) (éventuellement internationaux)
  - ENSAE Paris Tech
- Autre(s) partenaire(s) (éventuellement)
- Sites où la formation est dispensée (Etablissement, commune, département ou pays)
  - Ecole Polytechnique, Palaiseau, Essonne
  - Télécom ParisTech, Paris

II. CONTEXTE ET ENJEUX DE LA CREATION

- Justification du projet (évolution du secteur, évolution de la réglementation, secteur émergent scientifiquement etc.),

Le Big data marque le début d'une transformation majeure, qui va affecter de façon profonde l'ensemble des secteurs (de l'e-commerce à la recherche scientifique en passant par la finance et la santé).

L'exploitation de ces immenses masses de données nécessite des techniques mathématiques sophistiquées visant à extraire l'information pertinente. L'ensemble de ces méthodes forme le socle de la « science des données » (ou data science). Ce passage des données aux connaissances est porteur de nombreux défis qui requièrent une approche interdisciplinaire. La « science des données » s'appuie fortement sur le traitement statistique de l'information (statistiques mathématiques, statistiques numériques, apprentissage statistique ou *machine learning*). De l'analyse de données exploratoires aux techniques les plus sophistiquées d'inférence (*modèles graphiques hiérarchiques)* et de classification ou de régression *(deep learning, machine à vecteurs de support*), une vaste palette de méthodes de statistiques mathématiques et numériques et d'apprentissage est mobilisée. Ces méthodes, pour pouvoir être développées à l'échelle de masses de données requièrent la maitrise des mécanismes de distribution des données et des calculs à très grande échelle. Les mathématiques appliquées (analyse fonctionnelle, analyse numérique, optimisation convexe et non convexe) ont également un rôle essentiel à .jouer.

D'un point de vue applicatif, la « science des données » impacte fortement de nombreux secteurs. Il existe actuellement partout dans le monde un large déficit de "Data Scientists" et "Data Analysts". Les étudiants issus de formations en *science des données* et "*Big Data*" sont donc très attendus sur le marché de l'emploi.

A l'instar de tous les domaines d'innovations de ruptures (biotechnologies, e-médecine), le besoin d'ingénieurs de très haut-niveau et de doctorants est également important.

Le métier de data scientists se décline de nombreuses façons, allant de la mise en place de nouvelles générations de systèmes d'informations décisionnels, modifiant profondément la gestion des entreprises, aux développements d'applications complètement nouvelles (autour du e-commerce, de la recommandation, du minage de réseaux sociaux, fusion d'informations hétérogènes pour la finance – gestion- ou pour la santé).

- Place de la formation dans les contextes régional et national et éventuellement dans le cadre de la COMUE.

Cette formation a vocation à rejoindre la mention « Mathématiques et Applications » de l'Université Paris-Saclay (ouverture prévue à la rentrée 2015). Elle sera adaptée pour s'insérer dans l'offre de modules proposés dans ce parcours, l'objectif visé étant clairement la mutualisation des enseignements et la coopération entre les opérateurs de ce parcours.

La version qui sera déposée dans le cadre de la mention « Mathématiques et Applications » associera, en plus de l'Ecole Polytechnique et Télécom ParisTech, l'Université de Paris-Sud et l'ENSAE ParisTech. Cette adaptation impactera les équipes pédagogiques responsables des différents enseignements, mais n'altérera pas l'architecture globale de la formation, notamment l'équilibre entre l'offre de cours magistraux et les enseignements par projets et professionnalisants. L'insertion dans le parcours « Mathématiques et applications » enrichira de façon significative l'offre d'enseignements optionnels au second semestre, que nous avons volontairement limitée dans la version actuelle pour concentrer nos efforts sur le tronc commun.

Cette formation est complémentaire de différents parcours proposés dans la mention « informatique » de l'Université Paris-Saclay. Par rapport à ces parcours, l'accent est ici mis sur les méthodes mathématiques d'extraction de l'information. Ce Master recrutera des étudiants formés en mathématiques au niveau d'un M1 de mathématiques. A l'inverse, les pré-requis informatiques n'iront pas au-delà des connaissances « classiques » de la fin de la seconde année d'études dans des Ecoles d'ingénieurs (maitrise d'un langage de programmation orienté objet, une introduction aux bases de données et aux technologies du web)

- Adossement à la recherche (laboratoires, écoles doctorales liens avec cet environnement),

- Centre de Mathématiques appliquées (CMAP), Ecole Polytechnique
- Laboratoire d'Informatique (LIX), Ecole Polytechnique
- Laboratoire de Traitement et de Communication de l'Information, Télécom ParisTech
- Centre de Recherches en Economie et Statistique (ENSAE, Groupe des Ecoles Nationales d'Economie et de Statistique de l'INSEE)
- Département de Mathématiques de l'Université d'Orsay : équipe de probabilités et statistique (Université de Paris-Sud).

- Relations avec le milieu socioprofessionnel (entreprises partenaires, tissu industriel)
  - Aéronautique et Automobile (Peugeot-Citroen)
  - Banque (BNP Paribas, Société Générale, AXA),
  - Energie (GdFSuez, EDF, start-up du domaine)
  - Informatique (IBM, Google, start-up du domaine)
  - Télécommunications (Orange)
  - e-commerce ,
  - Média et industries des loisirs

III. OBJECTIFS DE LA FORMATION

- Objectifs en termes de connaissances scientifiques à acquérir (orientations scientifiques de la formation)
  - Statistique mathématique : régression linéaire et non-linéaire, modèles additifs généralisés, méthodes de bases de classification supervisée et non supervisée, tests d'hypothèses multiples, modèles graphiques.
  - Statistique en grandes dimensions et à large échelle: méthodes parcimonieuses (LASSO), régularisation, choix de modèles, aggrégation de modèles
  - Apprentissage statistique : méthodes à noyaux, apprentissage en ligne et distribué, ranking, recommandation, filtrage collaboratif, apprentissage par renforcement (bandits, processus de décision markoviens)
  - Graphes et inférence de graphes : modèles de graphes aléatoires, analyse statistique de graphes (détection de communauté), graphes dynamiques, signaux sur des graphes
  - Analyse de données textuelles : indexation, classification, catégorisation, analyse sémantique
  - Bases de données réparties : architecture des bases de données avancées, NoSQL (Hadoop, Map / Reduce), cloud, architecture matérielle et logicielles
  - Calcul distribué : analyse numérique matricielle pour très grandes matrices creuses, optimisation convexe et non convexe en grandes dimensions, infrastructure de calcul
- Objectifs en termes de compétences professionnelles à acquérir
  - Maitrise des infrastructures matérielles et des outils logiciels du domaine
  - Acquisition de données non structurées : recherche d'information, représentation des informations, crawling,
  - Manipulation des données : nettoyage, archivage, stockage (MySQL, NoSQL, bases colonnes, bases graphes), analyse de la qualité des données, intégration des données, recherche de métadonnées
  - Utilisation des réseaux sociaux et des moteurs de recherche
  - Traitement de données (séries temporelles, flux de données) et données imprécises
  - Visualisation des masses de données
  - Exploitation des données, décisionnel
  - Aspects légaux, éthiques et sociaux, confidentialité, anonymisation des données

- Objectifs en termes de débouchés (insertion professionnelle ou poursuite d'études) : secteurs principaux d'activités, profils de postes visés, métiers actuels ou futurs concernés

Plusieurs types d'entreprises/institutions sont à l'origine et/ou accompagnent aujourd'hui la tendance « Big Data ». Elles doivent recruter un nombre croissant de « data scientists ».

- **Les éditeurs de logiciels.** (e.g. MAPR, Pentaho, Teradata, VMware, IBM, Oracle, SAP), proposent aux entreprises des solutions compatibles avec la plate-forme d'analyse de données « open source » Hadoop. Selon les résultats d'un récent sondage publié par Capgemini, 58% des 600 directeurs du développement et cadres du secteur IT en France ont l'intention d'investir dans des systèmes Big Data comme Hadoop, au cours des trois prochaines années, un marché est donc en train de s'ouvrir. Les éditeurs commercialisent aujourd'hui des plates-formes de calcul pour le traitement de données massives collectées de manière séquentielle. L'intérêt de telles solutions pour les utilisateurs finaux dépend essentiellement de l'intégration dans ces plates-formes d'outils efficaces de type « informatique décisionnelle ». De tels outils d'aide à la décision, pour être performants, doivent reposer sur des algorithmes d'apprentissage « on-line » et « distribués » permettant de produire des résultats en « temps réel », suffisamment performants pour apporter une plus-value. Le succès de ces outils dépend de façon cruciale du développement de méthodes de type « machine-learning » dédiées, capables de supporter le « passage à l'échelle » sur les plates-formes de calcul existantes.

- **L'internet et l'e-commerce.** Les sociétés offrant, à partir des flux de données massives aujourd'hui disponibles, des services nouveaux, liés au web le plus souvent : moteur de recherche, de recommandation, ciblage, prédiction, etc. C'est le cas des géants du marché de l'internet (e.g. Google, Yahoo), mais aussi de sociétés émergentes à la croissance fulgurante telles que Criteo, proposant un ciblage comportemental des internautes à partir des « logs », en vue de la personnalisation de la publicité sur l'internet en fonction des centres d'intérêt ainsi identifiés. Dans le même ordre d'idée, Facebook a développé Beacon, un outil marketing analogue s'appliquant aux membres du fameux réseau social. Parmi les acteurs français proposant des services à partir des « Big Data », on pourrait également citer Exalead (moteur de recherche thématique), millemercis.com (moteur de recommandation), deezer.com (moteur de recherche « musical ») ou encore Twenga.fr (moteur d'achat). Ces acteurs, dans le secteur en pleine expansion de l'e-commerce en particulier, sont amenés à gérer/administrer des données à la fois massives et complexes (description d'objets à vendre, images, comportements de consommateur, etc.), et à développer des outils de type data mining pour en extraire de l'information (c'est le cas par exemple d'Amazon, de FNAC.com, ou de Sarenza).

- **Services Publics.** Dans le cadre du programme de modernisation des services de l'Etat, les ministères et organismes publics sont également amenés à collecter, administrer et analyser des volumes de données croissants. Les applications sont diverses, elles concernent en particulier le développement d'outils de surveillance, pour la Sécurité et la Défense, mais aussi pour la détection de fraudes à l'Assurance Maladie par exemple.

- **L'industrie High-Tech**. (On l'appellerait plus précisément Electronique Grand Public - Consumer Electronics en anglais - et Télécoms) Ces dernières années ont vu l'essor progressif des datasciences dans de nombreux secteurs industriels : vision par ordinateur (biométrie, contrôle d'accès, véhicules autonomes) reconnaissance vocale pour des interfaces homme-machine (smartphone), filtrage et aggrégation de contenus (filtrage de spam, gestion de flux web ou RSS), surveillance de systèmes complexes et critiques (aéronautique, énergie) etc.

- **Le secteur bancaire/financier.** Le caractère massif, temporel et haute-fréquence des données historiques financières/bancaires (cours, carnets d'ordre, information économique, mouvements de trésorerie, etc.) en fait un domaine d'application majeur de la « Science des Données », à des fins de gestion du risque de marché en Finance, du risque clientèle dans le domaine de la Banque de Détail.

Dans d'autres domaines, tels que **le secteur biomédical** avec l'avènement de la médecine individualisée (analyses post-génomique, métabonomique/métabolomique) ou encore celui de la **grande distribution**, des bases de données complexes et massives sont constituées, leur exploitation requérant des compétences avancées dans le domaine de l'analyse de données (CRM, bioinformatique).

- Objectifs en termes de flux pour la prochaine période.
    - Les formations de datascientistes sont très demandées ; il y a en particulier une forte demande pour des datascientistes « haut de gamme » combinant des compétences en gestion de données et en traitement statistique de l'information. L'objectif en termes de flux est de l'ordre de 25 à 40 étudiants

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

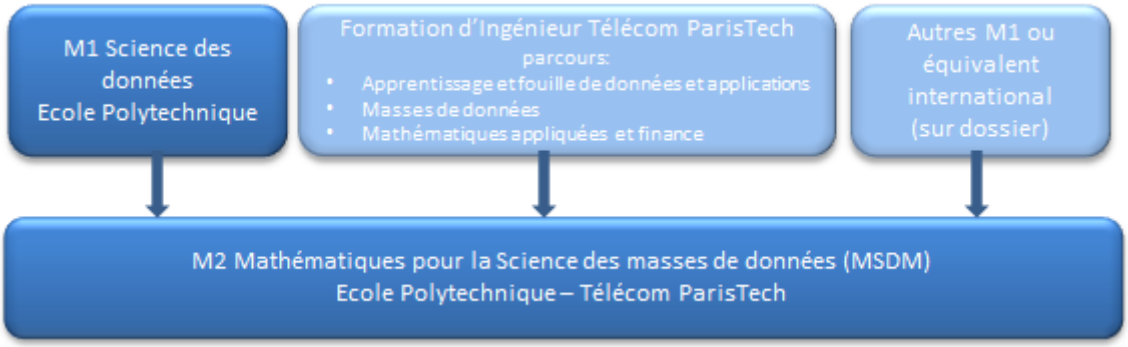
MINISTÈRE
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

Dossier de création de master

Partie B : Présentation détaillée de la formation

I. ORGANISATION

Le Master est structuré autour du M1 "Science des données" et du M2 "Data Sciences". Le M2 est ouvert aux étudiants issus d'un M1 ou équivalent international cohérent avec le domaine sur évaluation du dossier par le comité de pilotage.
L'accès au M2 pour les élèves de Télécom ParisTech se fait à l'issue de la deuxième année du cycle ingénieur.



-Structure du M1:
Le M1 est constitué de 3 cours scientifiques obligatoires (tronc commun) complétés par 5 cours au choix parmi une liste spécifique, des cours de formation humaine (Langues étrangères, humanités et sciences sociales) et un stage de 3 mois d'avril à juin.

-Structure du M2:
Le M2 est composé de 8 unités d'enseignement scientifique, d'un Projet de groupe tutoré (140h) et d'un stage

Le programme d'initiation à la recherche consiste en :

1. un groupe de lecture,
2. un séminaire.

Le groupe de lecture permet à la fois de confronter les étudiants à des problèmes de recherches actuels et de leur apprendre à effectuer un travail de recherche bibliographique (analyse et synthèse d'articles scientifiques). Après une phase de préparation de quelques semaines, chaque étudiant présente le travail réalisé et répond aux questions de l'auditoire à la fin de l'exposé.
Les séminaires visent à donner un aperçu des principaux sujets de recherche dans le domaine des data sciences ainsi que certains domaines connexes, et à présenter les enjeux économiques et sociétaux associés. Les exposés sont faits par des intervenants académiques et industriels n'appartenant pas, sauf

ponctuellement, à l'équipe pédagogique. Il est envisagé d'inviter régulièrement des chercheurs internationaux d'un haut niveau scientifique.

- **Stages**

La formation comporte 2 stages (15 ECTS et 20 ECTS respectivement), un en M1 d'une durée de 3 mois et l'autre en M2 d'une durée de 5 mois.
Le stage est un moment important dans la formation.. Il permet de confronter les enseignements théoriques reçus pendant la période des cours à la réalité d'un projet, et de mettre le participant en situation de réaliser et fournir un vrai travail de réflexion personnel sur un sujet en relation avec l'enseignement.

Une "bourse des stages" sera organisée fin décembre. Une convention de stage est établie entre l'élève, l'entreprise d'accueil et l'école qui précise les modalités, les dates et le sujet du stage de même que le nom du correspondant de stage de l'école et du responsable du stage dans l'entreprise. Le candidat choisit librement un stage proposé par l'un des enseignants du master, un stage en entreprise proposé dans le cadre de la "bourse des stages", ou un stage d'origine différente ayant reçu l'agrément d'un enseignant du master.

À l'issue de son stage, un étudiant doit :

- Remettre un rapport scientifique et technique décrivant le travail effectué et les résultats obtenus. Ce rapport est évalué par le tuteur de stage. Il doit permettre à un jury d'avoir les éléments nécessaires et suffisants pour statuer sur la qualité du travail fourni.
- Effectuer une présentation orale (« soutenance ») devant un jury présidé par le tuteur de stage. La durée normale pour la soutenance est d'une heure au total y compris les questions, les démonstrations et la délibération du jury.

- **Dispositifs d'acquisition des compétences pré-professionnelles**

Ce parcours associe des cours théoriques et méthodologiques complétés par des projets en « vrai grandeur » faisant intervenir tous les aspects des sciences des données, depuis l'acquisition jusqu'à l'exploitation et l'analyse. Une des originalités de ce parcours sera un recours à des pédagogies innovantes basées sur l'apprentissage par projets. Le module 1 (projet tutoré, 140h) vise à confronter les acquisitions théoriques et la pratique, à s'approprier les problématiques du « big data » à travers un exemple en « grandeur nature » et en interaction avec des professionnels (et des tuteurs académiques). Les participants travaillent par groupes de 5 à 6 étudiants sur un cas réel sous la responsabilité d'un encadrant professionnel et d'un enseignant du master qui jouent le rôle de "chef de projet" et de clients. Ce projet permettra aux étudiants d'interagir avec des partenaires industriels des "big data" qui feront bénéficier les étudiants de leur expérience acquise dans le secteur. Il a aussi pour objet de développer de travail en équipe, la gestion d'un projet, la définition et la rédaction d'un cahier des charges, la mise en œuvre informatique du projet, et la présentation de livrables et d'un produit final.

- **Publics concernés (filières de recrutement des étudiants, les modalités de recrutement à l'entrée de la mention, en M1 et en M2)**

Le recrutement en M1 se fait sur dossier, par le jury d'admission. Les principales filières de recrutement sont les élèves de l'Ecole Polytechnique, de Télécom ParisTech, de l'ENSAE ParisTech ainsi que les étudiants de l'Université de Paris-Sud et de grandes universités étrangères.

Le recrutement en M2 se fait :

- pour les étudiants du M1 après validation de 60 ECTS.
- pour les autres étudiants, après validation de 60 ECTS d'un M1 français ou étranger dont le programme est jugé équivalent par le comité de pilotage.

➢ **Contrôle des connaissances et des compétences**

Chaque activité est assortie d'une note sur 20. Une note minimale de 10 est exigée pour obtenir l'ensemble des ECTS associés. La note est basée :

- pour chaque unité d'enseignement, sur un contrôle de connaissance écrit ;
- pour le projet sur une réalisation logicielle et une soutenance ;
- pour le groupe de lecture, sur la présentation d'un article ;
- pour le stage, sur un mémoire et une soutenance.

➢ **Pilotage de la formation**

Le comité de pilotage (CP) est composé des responsables de la formation de chaque entité partenaire et d'une sélection d'enseignants et d'industriels. Le comité de pilotage veille au bon fonctionnement général de la formation. Il assure l'organisation des cours, la sélection et l'admission des futurs élèves. En outre, il assiste les élèves dans leur recherche de stage et la validation des sujets proposés.

L'évaluation et l'évolution de l'enseignement est complétée par une évaluation des enseignements effectuée par voie de questionnaires distribués aux étudiants et réalisée à la fin de chaque période d'enseignement. Elle sera accompagnée d'un temps d'échange oral avec les étudiants ou les représentants qu'ils auront désignés à cet effet. Une synthèse des résultats de ces évaluations sera présentée au comité de pilotage, qui les utilisera dans ses propositions d'évolution de la formation, en coordination avec les directions de chaque établissement chargées de l'enseignement au niveau master.

Les diplômés de ce master seront suivis au même titre que les diplômés des autres formations d'établissements partenaires. Ils seront intégrés aux annuaires des anciens élèves et participeront aux enquêtes métiers permettant le suivi de carrières des diplômés.

II. DESCRIPTIF DE LA FORMATION

La formation permet d'acquérir 120 ECTS dont 60 en M1 et 60 en M2, la répartition de ces crédits est indiquées dans le tableau ci-après :

	Unités d'enseignement techniques		Formation humaine	Stage	TOTAL
	Tronc commun	Options			
M1	15 ECTS	25 ECTS	5 ECTS	15 ECTS	60 ECTS
M2	40 ECTS			20 ECTS	60 ECTS

- **Descriptif des UE du M1**

Tronc commun (3 UE, 15 ECTS):

1. **Apprentissage statistique et estimation non paramétrique (40h, 5 ECTS) Ecole Polytechnique**

Le but de ce cours est de donner une introduction mathématique à la Théorie de l'apprentissage statistique et à l'Estimation Non-paramétrique. Nous allons également présenter quelques algorithmes principaux de réalisation.

2. **Bases de données et gestion de Bigdata (40h, 5 ECTS) Ecole Polytechnique**

Ce cours présente d'une part la gestion des bases de données traditionnelles et d'autre part introduira les éléments et les techniques propres aux masses de données (Big Data).

En ce qui concerne les bases de données traditionnelles le cours comprendra : modèle relationnel, SQL, conception de base de données, indexation, traitement des requêtes et optimisation Web - interfaces de base de données et des services Web. La partie Bigdata du cours présentera la technologie Bigdata (Hadoop, MapReduce, Pig, Hive, des solutions de stockage NoSQL focalisées sur Hbase), les algorithmes

d'apprentissage automatique distribué et l'analyse de données Bigdata avec des applications dans la publicité sur le Web, la théorie des graphes et la fouille de données.

3. **Big Data management : data mining (40h, 5 ECTS) Ecole Polytechnique**

Ce cours s'adresse à tous les scientifiques désireux de comprendre les méthodes de recherche d'information dans de grands ensembles de documents, de donner des moyens de caractériser des ensembles à partir d'exemples et de retrouver dans des données celles qui s'approchent le plus de motifs précalculés. Le traitement de ces données utilise les méthodes classiques de programmation, mais aussi fait appel à de nouveaux concepts comme l'apprentissage et la fouille de données (supervised learning, datamining, text mining, retrieval, etc.).

Cours aux choix (5 UE; 25 ECTS):

1. **Recherche opérationnelle : aspects mathématiques et applications (40h, 5 ECTS) Ecole Polytechnique**

Le cours présente quelques grandes familles de méthodes mathématiques utiles en recherche opérationnelle, afin de donner la capacité de modélisation, et de permettre de reconnaître les problèmes pour lesquels des algorithmes rapides de résolution existent. On met l'accent sur les techniques issues de la programmation linéaire ou convexe, qui sont souvent à l'origine de tels algorithmes.

2. **Traitement du Signal (40h, 5 ECTS) Ecole Polytechnique**

Ce cours propose une introduction au traitement du signal. Il requiert un savoir-faire de base en analyse (transformée de Fourier) et en probabilités (variables et processus aléatoires). Le cours débute par une présentation du filtrage analogique qui met en avant le rôle central de l'analyse de Fourier dans ce domaine. L'essentiel du traitement du signal étant devenu numérique, il se poursuit par une étude, à l'aide du même outil, de la conversion analogique/numérique et de l'importance de l'algorithme de Transformée de Fourier Rapide (FFT en anglais) dans le filtrage numérique.

3. **Utilisation de l'aléatoire en algorithmique (40h, 5 ECTS) Ecole Polytechnique**

Il y a plusieurs manière d'utiliser l'aléatoire en informatique. On peut étudier le comportement d'un processus informatique (programme, protocole, automate, …) face à un modèle aléatoire, ou encore construire des générateurs d'instances aléatoires (séquences, arbres, graphes, …) à des fins applicatives. Une autre approche, celle de ce cours, considère l'aléatoire comme une nouvelle ressource permettant de résoudre plus efficacement certains problèmes. Par exemple, on ne sait pas générer un nombre premier sans tirer au sort, ou encore la congestion dans les réseaux ne peut pas être évitée sans un protocole de routage en partie aléatoire.
De manière contre-intuitive, prendre des décisions au hasard permet de concevoir des algorithmes plus simples et souvent plus efficaces que leurs analogues déterministes. Les domaines et les applications sont vastes. Ce cours fournit une présentation accessible à la plupart de ces derniers et de leurs idées centrales. Chacune de ces idées sera présentée à travers des exemples simples et motivés.

4. **Analyse d'Images et Vision par Ordinateur : algorithmes et applications (40h, 5 ECTS) Ecole Polytechnique**

L'objectif de la vision par ordinateur est de calculer les propriétés du monde réel à partir d'images numériques. Les problèmes abordés comprennent l'identification de la forme 3D d'un environnement, l'estimation du mouvement et la reconnaissance de personnes et d'objets, le tout à travers l'analyse d'images et de vidéos.

Ce cours est une introduction à l'analyse d'image et à la vision par ordinateur au travers de sujets tels que la détection de caractéristiques, la segmentation d'images, l'estimation du mouvement, les mosaïques d'images, la reconstruction de forme 3D et la reconnaissance d'objets. Ces sujets seront abordés sous l'angle des algorithmes et des outils mathématiques. Les applications seront développées en C++. La connaissance de ce langage n'est pas un prérequis et une partie des cours sera consacrée à son apprentissage.

5. **Algorithmique avancée (40h, 5 ECTS) Ecole Polytechnique**

Ce cours porte sur des techniques avancées dans la conception et l'analyse des algorithmes. Le cours commence par un parcours rapide de quelques uns des paradigmes principaux de l'algorithmique, notamment flots et couplages et programmation linéaire. Il revient ensuite rapidement sur la NP-complétude et notamment sur l'importance des réductions polynomiales. Au delà de l'analyse de pire cas, on envisage plusieurs approches possibles de l'analyse d'algorithmes, d'une part au travers des notions de complexités: pire-cas, en moyenne, amortie, lissée, ou paramétrique, et d'autre part au travers de mesures de qualités de sortie: optimalité, facteur d'approximation pour l'optimisation, de compétitivité pour l'algorithmique on-line. Ce sera l'occasion de présenter entre autres les notions de potentiel, de noyau, de largeur arborescente

6. **Méthodes de Monte-Carlo et processus stochastiques: du linéaire au non-linéaire (40h, 5 ECTS) Ecole Polytechnique**

Un enjeu fondamental du programme de Mathématiques Appliquées est de modéliser et simuler des systèmes complexes pour comprendre leur comportement qualitatif et quantitatif. Ce cours introduit des méthodes probabilistes effectives de calcul et de simulation, principalement axées sur les processus à temps continu, avec dynamique linéaire puis non-linéaire ("interactions entre particules"). Un souci permanent est leur validation, leur efficacité numérique et leur illustration dans les situations concrètes, tirées notamment de l'ingénierie financière, l'écologie évolutive, les réseaux de communication, la mécanique des fluides, la physique et la chimie, entre autres… Ces méthodes ont pris une importance déterminante dans des domaines applicatifs stratégiques variés

7. **Analyse des séries temporelles (40h, 5 ECTS) Ecole Polytechnique**

Ce cours introduit les méthodes d'analyse, d'estimation statistique et de prédiction des séries chronologiques. Ces types de traitement statistique sont utilisés dans de nombreux domaines, des sciences de l'ingénieur aux sciences sociales, tels que l'économétrie, l'hydrologie, les problèmes de détection, localisation et poursuite de cibles, la métrologie réseau, etc. Le cadre des propriétés du second ordre et de la prédiction linéaire permettra dans un premier temps une description détaillée des approches statistiques les plus répandues d'analyse et de traitement de données temporellement dépendantes. Ces méthodes d'analyse temporelle ou spectrale s'appliquent à une vaste classe de modèles stationnaires. Le cas des modèles espace-état linéaires seront ensuite étudiés. Des algorithmes numériques pour la prédiction et l'inférence statistique seront obtenus dans ce contexte. Cette approche de la modélisation sera étendue à l'étude de modèles espace-état non-linéaires. Ceci permettra de prendre en compte différents types de dépendances temporelles en vue de traitements statistiques efficaces. Enfin, nous consacrerons un cours au principe d'invariance d'échelle et son utilisation pour tester des hypothèses statistiques.

8. **Randomization in Computer Science: Games, Networks, Epidemic and Evolutionary Algorithms (40h, 5 ECTS) Ecole Polytechnique**

Randomized methods are one of the key tools in computer science, clearly not limited to their best-known use in randomized algorithms. Our aims for this course are twofold: (i) We shall give

an introduction to the diverse applications of randomized methods in computer science. (ii) Parallel to this, we shall develop a small, but powerful set of mathematical tools that suffice to understand most uses of randomness in computer science.

9. **Programming C++ (40h, 5 ECTS) Ecole Polytechnique**

Le but de ce cours est de donner une connaissance pratique de C++. L'environnement de travail est librement choisi. Les élèves peuvent utiliser leur ordinateur portable. Un environnement de travail avancé est souhaité (e.g. Visual Studio, Xcode, Eclipse, etc.).
Les sujets étudiés sont par exemple la mémoire et les pointeurs, le polymorphisme de fonctions et les opérateurs, la programmation orientée objet, l'héritage, les patrons (templates), etc. Ce cours suppose la connaissance d'au moins un autre langage de programmation générique (Java, C, Fortran, Visual Basic ...).

- **Descriptif des UE du M2:**

- **Projet de groupe "science des données" (Projet de groupe tutoré 120h ; 10 ECTS), Télécom ParisTech, Ecole Polytechnique; partenaires industriels**

L'objet de ce module est de familiariser les étudiants avec la problématique du Big Data. L'idée de ce module est de partir de problèmes "big data" et de développer des solutions complètes, allant de la prise en compte du problème, du développement de méthodes de traitement, jusqu'à la mise en oeuvre algorithmique. Les méthodes de statistique exploratoire et d'apprentissage utilisés sont élémentaires dans leurs principes (régression linéaire, régression logistique, naïve Bayes, *k*-plus proches voisins, analyse discriminante, …), la difficulté est ici de comprendre comment ces méthodes s'appliquent sur des masses de données et de les mettre en oeuvre sur des masses de données réparties en initiant les étudiants aux mécanismes de base des BigData (Hadoop, MapReduce). Ce module doit permettre aux étudiants d'acquérir une compréhension fine des problèmes de modélisation, de diagnostic des modèles, de nettoyage des données, et bien sûr le développement d'algorithmes efficaces permettant le passage à l'échelle et le traitement à l'échelle du "big data". Il introduit également des technologies émergentes dans le domaine de la gestion de données hétérogènes, massives, complexes ou semi-structurées. Il présente en particulier les grands principes et la typologie des systèmes NoSQL ainsi que quelques implémentations.

- **Statistique en grande dimension (CM 40h, 5 ECTS) : Ecole Polytechnique ; Télécom ParisTech**

Les données de très-grandes dimensions sont aujourd'hui la règle plutôt que l'exception. L'explosion combinatoire (« curse of dimensionality ») rend nécessaire d'utiliser des méthodes spécialement adaptées aux problèmes d'apprentissage où les observations dépendent d'un nombre très important de variables explicatives. Ce module présente à la fois les aspects méthodologiques, les algorithmes, et les applications de l'inférence en grande dimension.

- **Apprentissage statistique avancé (CM 40h, 5 ECTS) : Télécom ParisTech, Ecole Polytechnique**

Le module donne un aperçu des méthodes avancées de l'apprentissage statistique. Le cours traitera tout d'abord des machines à vecteur de support en partant de la théorie des espaces à noyaux auto-reproduisants puis en introduisant les applications de ces espaces pour construire des méthodes de régression et de classification (noyau, classification à larges marges, "kernel trick"). Le cours présentera ensuite une introduction aux réseaux de neurones (neurones multi-couches) en insistant sur les avancées les plus récentes dans ce domaine (apprentissage profond, deep-learning) et leurs applications à différents domaines. Il introduira succinctement les méthodes ensemblistes dont le "bagging" et le "boosting" mais aussi les techniques d'agrégation de prédicteurs (approche en ligne et hors ligne); un intérêt particulier sera porté aux applications des méthodes ensemblistes aux arbres de classification et de régression. Il se

conclura par une brève introduction à la problématique de l'apprentissage par renforcement (bandits multi-bras, processus de décision markoviens, Q-learning, TD-learning)

- **Architecture et calcul distribué (CM 40h, 5 ECTS), Télécom ParisTech, Ecole Polytechnique**

Ce cours est une introduction aux architectures, et environnements de calcul parallèle et distribué, et un apprentissage théorique et appliqué de leur algorithmique et programmation. Ce cours introduira des notions d'architecture, des bases d'algorithmiques, des langages de programmation parallèle et distribuée. Des développements plus spécifiques à l'analyse numérique matricielle (résolution et analyse spectrale de grands systèmes creux) et à la résolution de problèmes d'optimisation convexes en grandes dimensions seront également introduites.

- **Modèles graphiques (CM 40h, 5 ECTS) : Télécom Paris Tech**

Ce module introduit les concepts fondamentaux des modèles graphiques hiérarchiques et les algorithmes d'inférence exacte ou approchée associés. Les modèles graphiques sont utilisés dans de nombreux domaines de l'apprentissage statistique et sont à la base de nombreuses méthodes développées pour résoudre des problèmes de modélisation et d'inférence de données complexes. Le cours introduit tout d'abord les modèles graphiques sous un angle probabiliste, comme modèle de structure de dépendance complexe (graphes non dirigés, réseaux markoviens). Il présente ensuite l'inférence et l'apprentissage des modèles statistiques (inférence exacte, inférence approchée, approches variationnelles). Ce module introduit à cette occasion les méthodes de simulation par Méthodes de Monte Carlo par Chaînes de Markov et les méthodes de simulation par systèmes de particules en interaction.

- **Données textuelles (CM 20h, 2,5 ECTS) : Ecole Polytechnique, Telecom ParisTech**

Les données textuelles représentent un enjeu très important dans de nombreuses applications (web, réseaux sociaux, marketing, enquêtes et études d'opinion, etc.) L'objectif de ce module est de présenter les principales caractéristiques de ce type de données ainsi que les modèles et méthodes permettant d'indexer, de catégoriser et de rechercher l'information présente dans des données textuelles. Les études de cas associées au cours porteront en particulier sur des applications dans le domaine des services web et des réseaux sociaux pour des tâches telles que la personnalisation et la recommandation de contenu, le filtrage de documents, l'analyse d'opinions et le marketing en ligne.

- **Réseaux, graphes, (CM 20h, 2.5 ECTS), Ecole Polytechnique, Télécom ParisTech**

Les graphes sont au coeur de nombreuses problématiques (données du web, réseaux sociaux, signaux sur les graphes). La science des réseaux (network science) est un nouveau domaine interdisciplinaire s'intéressant aux graphes des réseaux complexes et à leurs applications. Ce cours a pour objectif de présenter les fondements mathématiques des méthodes de la théorie des réseaux et des graphes ainsi que leurs applications à différents types de réseaux. Ce cours décrit l'état de l'art et ses applications sur des réseaux (web, sociaux, contacts, etc…), incluant les méthodes et les outils de pré-traitement des graphes, jusqu'à la recherche, le classement et l'évaluation de communautés d'utilisateurs

- **Visualisation analytique (CM 20h, 2.5 ECTS)**

Le cours a pour but d'introduire les concepts et techniques relatifs à la visualisation d'information. Le domaine de visualisation d'information concerne des techniques de représentation de données complexes, n'ayant pas une représentation naturelle et évidente. Il est situé à l'intersection de graphisme, de l'interaction humain-machine (IHM) et des sciences cognitives. Il a pour objectif, en proposant une présentation appropriée à différents types de structures, d'obtenir une meilleure compréhension des données abstraites et complexes, telles que des données symboliques, tabulaires, hiérarchiques, textuelles ou en réseau. Le cours approfondira les méthodes pour visualiser et naviguer dans des masses de données. Il présentera aussi les outils logiciels pour réaliser et déployer des systèmes d'analyse visuelle pouvant gérer, chercher, visualiser et analyser des masses de données.