Programme des enseignements

3e année

Filière Statistique et ingénierie des données

ANNÉE SCOLAIRE 2019 / 2020

ENSAI

École nationale
de la statistique
et de l'analyse
de l'information

FILIÈRE STATISTIQUE ET INGÉNIERIE DES DONNÉES

ANNÉE SCOLAIRE 2019/2020

Data Science and Big Data

2019/2020 ACADEMIC YEAR

Table des matières

Présentation de la filière

L'objectif de la filière Statistique et Ingénierie des Données est d'offrir aux étudiants une double compétence recherchée sur le marché actuel du travail où il est impossible de se passer d'informatique et où le traitement de données de plus en plus volumineuses nécessite de sérieuses connaissances statistiques. Au cours de cette 3ème année, les élèves vont acquérir des bases informatiques solides qui leur permettront de maîtriser la conception de systèmes. Ils pourront ainsi mieux définir leurs attentes afin d'exploiter, à l'aide d'outils statistiques, les nouvelles mines d'information disponibles par le monde. La filière SID ne fixe pas de domaine d'application, celui-ci peut être défini pendant l'année voire pendant le stage, cette filière apporte essentiellement des compétences complémentaires en informatique, elle garde ouverte les portes de la finance, de la biostat, du marketing ou tout autre domaine et permet bien entendu de poursuivre en thèse. Cette filière forme aux métiers de *data scientists* et *data analyst*.

Au cours des deux premières années à l'ENSAI, les étudiants ont acquis les bases de la conception orientée objet (Java), ils ont appris à utiliser des outils de développement spécialement adaptés (Eclipse par exemple), à modéliser des applications (UML) et à concevoir des bases de données. Ils ont également assimilé les principales méthodes statistiques. Toutes ces connaissances ont été appliquées et consolidées au cours de projets en deuxième année.

S'appuyant sur ces acquis, la filière s'articule autour de deux axes principaux : l'outil informatique d'une part, l'apprentissage statistique et l'ingénierie des données d'autre part.

Les outils informatiques
Le développement de grandes applications informatiques nécessite d'utiliser des méthodes d'aide à la conception. Les méthodes de développement d'application sont présentées dans le cours de Génie logiciel. Les architectures des grandes applications sont désormais de type multi-tiers, pour les appréhender, il est nécessaire d'avoir des connaissances de base en architecture des systèmes, en réseaux, en sécurité. Le développement d'applications est réalisé sur la plate-forme J2EE. Les technologies du Web Sémantique seront également abordées.

L'ingénierie des données
Le rôle du statisticien – informaticien est d'analyser des données à l'aide de méthodes statistiques et de l'outil informatique. Nous pouvons définir 4 phases dans le traitement informatique des données : la récupération, le stockage, l'analyse et la visualisation des résultats. Ils apprendront à manipuler de très grands volumes de données, à créer des entrepôts de données et à effectuer une analyse multi-dimensionnelle de ces données. Les cours d'apprentissage statistique leur permettront d'extraire de la connaissance à partir de ces données. Ils découvriront comment inventer de nouvelles technologies de stockage et de gestion des données, dans le cadre du Big Data. ils aborderont également les problèmes de sécurité associés au traitement des données.

Plusieurs projets sont réalisés au cours de l'année afin de mettre en pratique les divers enseignements dispensés au cours de cette année. Ces projets sont réalisés individuellement ou en groupe, permettant aux étudiants de vivre la réalité d'un développement d'application.
Les projets sont de nature très diverses et peuvent être réalisés en partenariat avec des industriels ou des chercheurs. Ils ont tous pour but de mettre les étudiants en situation de statisticien ayant de bonnes compétences en informatique.

Volume horaire						
	Cours	**Ateliers**	**Projets**	**Total**	**Crédits**	**Enseignant-e-s**
UE0 Tronc commun						
Droit du Travail	3	6		9	0.5	Charlotte GRUNDMAN
Anglais	30			30	1.5	
Sport		30		30	0	
Total	**33**	**36**	**0**	**69**	**2**	
UE1 Machine learning						
Machine learning	18	21		39	3	Hong-Phuong DANG, Romaric GAUDEL, Fabien NAVARRO, Brigitte GELEIN
Machine-learning – Réseaux de neurones avancés	3	9		12	0	Romaric GAUDEL
Machine learning – Systèmes de recommandation	6	6		12	0	Romaric GAUDEL
Régression pénalisée et sélection de modèles	9	6		15	1	
Apprentissage statistique à grande échelle	9	9		18	1.5	Arthur KATOSSKY, Rémi PÉPIN
Webmining et traitement du langage	9	12		21	1.5	Arthur KATOSSKY
Total	**54**	**63**	**0**	**117**	**7**	
UE2 Développement d'application						
Génie logiciel	39	39		78	4	Mathieu ACHER, Johann BOURCIER, Olivier BARAIS, Benoît Combemale, Mohamed GRAIET
Indexation web	9	6		15	1	Nawfal TACHFINE
Projet Web - Applications web	12	15	9	36	3	Olivier BARAIS
Total	**60**	**60**	**9**	**129**	**8**	
UE3 Big Data						
Technologies Sémantiques	6	9		15	1	Sébastien FERRÉ
Technologies NoSQL	12	3		15	1	David GROSS-AMBLARD
Publication de données respectueuse de la vie privée	15	6		21	1	Tristan ALLARD
Total	**33**	**18**	**0**	**51**	**3**	
UE4 Systèmes et Réseaux						
Réseaux et systèmes d'exploitation	15	6		21	2	Jean-Baptiste LOISEL
Initiation à Unix	9	6		15	0	François-Xavier BRU
Systèmes Répartis	15	6		21	1	Davide FREY, George GIAKKOUPIS
Sécurité des données	9	6		15	1	Franck LANDELLE
Grandes masses de données sur Cloud	12	12		24	2	Gabriel ANTONIU
Total	**60**	**36**	**0**	**96**	**6**	
Projet de fin d'étude						
Projet méthodologique		3	9	12	1	
Projet de fin d'étude		9	27	36	3	
Data Challenge		12		12	0	
Total	**0**	**24**	**36**	**60**	**4**	
Séminaire professionnel						
Séminaire professionnel	30			30	0	
Total	**30**	**0**	**0**	**30**	**0**	
TOTAL	**270**	**237**	**45**	**552**	**30**	

UE Stage	**Crédits 25**

Descriptifs des enseignements communs

UE0 – Tronc commun

Droit du Travail

Work Law

Cours : 3h • Atelier : 6h

Enseignant : Charlotte GRUNDMAN, Avocat au Barreau de Paris
Correspondant : Ronan LE SAOUT

Objectif pédagogique :
La matière étant extrêmement vaste et complexe, il est ici proposé aux étudiants une approche didactique et vivante du sujet, l'objectif de l'enseignement étant de permettre aux étudiants qui travailleront dans un futur proche en entreprise d'avoir compris certaines notions pratiques essentielles en droit du travail.

Contenu de la matière

A cette fin, et hormis le cours d'amphi, il sera systématiquement proposé aux étudiants, après l'étude d'une notion, un exercice visant à mettre en pratique la notion abordée.

Afin de satisfaire le plus possible à cet objectif, il est ainsi proposé l'organisation suivante des cours :

Cours commun (3 heures) :

Chapitre 1 : Comprendre d'où l'on vient pour savoir où on va :

- Introduction historique au droit du travail

- Les sources du droit du travail
  - sources imposées,
  - sources négociées
- Ordre public absolu et ordre public social

Chapitre 2 : les instances de contrôle du droit du travail

- L'inspecteur du travail
- Les multiples juges du droit du travail
- Point sur la procédure prud'homale

Chapitre 3 : Formation et exécution du contrat de travail

- la qualification du contrat de travail : « faux artisans, faux auto-entrepreneurs et vrai salarié ».
- le contrat à durée indéterminée, norme juridique et sociale
- la période d'essai après la loi du 25 juin 2008 : définition, durée et rupture
- les principales clauses du contrat de travail :
- la clause de mobilité
- la clause de non-concurrence

Chapitre 4 : la rupture du contrat à durée indéterminée

- le licenciement pour motif personnel
- le licenciement pour motif économique
- la démission du salarié
- les autres modes de rupture

Les TD :

La première heure de cours sera consacrée à l'étude d'un chapitre. Cet exposé sera suivi d'une mise en situation pratique, où les étudiants devront par groupe répondre à un cas pratique. Un rapporteur sera désigné par groupe, et la notation se fera à cette occasion.

Chapitre 1 : La modification du contrat de travail

Modification du contrat de travail et changement des conditions de travail

- la durée du travail (focus sur le forfait-jour)
- le lieu de travail
- la rémunération

Chapitre 2 : Le recrutement

Chapitre 3 : les droits fondamentaux du salarié

- Le fait religieux en entreprise
- Vie personnelle et technologies de l'information et de la communication (TIC)
- La mise en place de moyens de contrôle via les TIC en entreprise
- Harcèlements
- Maladie et maternité du salarié

Langue d'enseignement

Français

UE0 – Tronc commun

Anglais

English

Cours : 30h (dont 15h d'aide au projet)

Enseignant : Divers intervenants
Correspondant : Todd DONAHUE

Objectif pédagogique

Les élèves qui n'ont pas passé ou qui n'ont pas réussi le TOEIC l'année dernière auront progressé dans les compétences requises – c'est à dire, la compréhension orale, la reconnaissance des erreurs, les pièges grammaticaux, et la compréhension écrite. Les autres auront acquis les compétences nécessaires pour affronter le monde professionnel. Ils auront vu les tournures qui aident à diriger et à participer à des réunions, à prendre des décisions, et à négocier. Ils se seront entraînés à faire des présentations. Ils auront rédigé un projet en anglais et préparé la soutenance de ce projet.

Contenu de la matière

Pour les élèves qui n'ont pas eu un score d'au moins 785 au TOEIC : pendant les 5 premières séances, la plupart des cours seront basés sur la préparation à cet examen. Les ressources informatiques de l'Ecole doivent aussi être mises à profit (pages Moodle, TOEIC Mastery), ainsi que les méthodes disponibles à la bibliothèque.
Pour les autres élèves, les cours seront organisés par groupe de niveau et conçus afin de les préparer à affronter le monde professionnel sur le plan international. Les thèmes suivants seront traités : « Leading meetings », « Interviews », « Presentations », « Taking decisions », et « Negociating deals », et « Cultural and Political Current Events ».

Ensuite, les 5 dernières séances seront consacrées au travail de rédaction/correction des rapports faits en anglais dans chaque filière ainsi qu'à la préparation des soutenances orales. Chaque responsable de filière indiquera aux élèves, en début d'année, le projet concerné et les modalités de notation. Les élèves recevront des consignes détaillées avant de démarrer ces cinq séances, afin d'arriver à la première séance avec une première version ou extrait de leur rapport en anglais prêt pour correction et relecture.

Pré-requis

Aucun

Contrôle des connaissances

L'examen final prend la forme d'une simulation d'entretien d'embauche. Cet examen oral durera environ 25 minutes, sera noté, et permettra d'évaluer le niveau d'expression orale sur l'échelle CECRL*. Le CV et la lettre faite pour cet exercice seront évalués et feront partie de la note finale. Le niveau acquis apparaîtra sur le Supplément au diplôme. L'objectif de la CTI† pour tous les élèves ingénieurs est d'atteindre le niveau B2.

* le **C**adre **e**uropéen **c**ommun de **r**éférence pour les **l**angues.
† la **C**ommission des **T**itres d'**I**ngénieur.

Références bibliographiques

- Arbogast, B., *30 Days to the TOEIC Test*, Canada: Peterson's, 2002.
- Schrampfer-Azar, B., *Understanding and Using English Grammar*, New York: Longman, 1999.
- Buckwalter, Elvis, et.al, *Boostez votre score au TOEIC-spécial étudiants*, Paris: Eyrolles, 2009.

- Gear, Jolene, *Cambridge Grammar and Vocabulary for the TOEIC Test*, Cambridge: Cambridge University Press, 2010.
- Lecomte, Stéphane, et. al, *La Grammaire au TOEIC et au TOEFL : Mode d'emploi*, Paris: Ophrys, 2008.
- Lougheed, Lin, *Tests complets pour le nouveau TOEIC* (4ème ed.), Paris: Pearson Education France, 2008.
- MBA Center, *New TOEIC Study Book*, Paris: MBA Center Publications, 2007.

Langue d'enseignement

Anglais

Pour tout complément d'information, chaque élève peut consulter le *Programme des enseignements : Langues étrangères*, distribué au début de l'année académique.

UE0 – Tronc commun

Sport

Sport

Atelier : 30h

Enseignant : Divers intervenants
Correspondant : Julien LEPAGE

Cours facultatif

Objectif de la matière

L'objectif est d'amener les élèves à maintenir un esprit sportif, sortir du strict cadre académique et développer leurs capacités physiques.

Contenu de la matière

9 activités sportives sont proposées par l'école :

- Badminton
- Basket
- Football
- Hand-ball
- Tennis de table
- Tennis débutant
- Volley-ball
- Cross-training
- Course à pied/préparation physique/coaching sportif

Outre les entraînements, les élèves inscrits peuvent être amenés à participer à des compétitions.

Prise en compte dans la scolarité

La participation à une activité sportive peut donner lieu à l'attribution d'un bonus ajouté sur la moyenne du semestre concerné. Le niveau de ce bonus est précisé dans une circulaire d'application en début d'année académique. Il varie selon l'assiduité aux séances, l'engagement et la participation aux compétitions tout au long de l'année.
Pour être définitive, la liste des élèves bénéficiant de ces bonus doit être validée par le directeur des études.

Un bonus peut être exceptionnellement attribué en dehors des activités sportives réalisées dans le cadre Ensai. Pour y prétendre, les élèves concernés doivent remplir les 3 conditions suivantes :
- pratiquer régulièrement une activité sportive et participer aux compétitions liées ;
- posséder un niveau national (voir très bon niveau régional suivant le sport en question) ;
- déposer une demande argumentée auprès de la direction des études et du service sport en début d'année scolaire, afin de faire valider le programme d'entraînement, des compétitions et les modalités de diffusion des performances.

Pour certains ayant des contraintes sportives, des aménagements horaires pourront d'ailleurs être ainsi envisagés si besoin.

UE1 – Machine Learning

Machine Learning

Machine Learning

Cours : 18h • Atelier : 21h

Enseignants : Hong-Phuong DANG (Ensai), Romaric GAUDEL (Ensai), Fabien NAVARRO (Ensai) et Brigitte GELEIN (Ensai)
Correspondant : Arthur KATOSSKY (Ensai)

Objectif pédagogique :

Ce cours présente les principes de l'apprentissage automatique (Machine Learning) ainsi que les modèles les plus utilisés.

Contenu de la matière

- Principes de l'apprentissage automatique
  - Apprentissage supervisé vs. non-supervisé ; échantillon d'entraînement et de validation, overfitting, erreur de généralisation ; fonction de coût (loss function) et minimisation d'une erreur ; évaluation des méthodes non-supervisées ; méthodes vues en 2A en tant que méthodes d'apprentissage
- Réseaux de neurones
  - Principe des réseaux de neurones ; propriétés des réseaux de neurones simples ; descente de gradient ; réseaux de neurones profonds ; architectures particulières (ex: réseaux de convolution ; réseaux récurrents ; ...) ; réduction de la dimension à l'aide de réseaux de neurones (auto-encodeurs ; word2vec ; ...).
- Méthodes d'agrégation
  - Quelques rappels et approfondissements (CART, multiregression trees), Bagging, random forests, Boosting, XGBoost, Stacking (agrégation de modèles de types différents par construction d'un modèle « superviseur » qui combine au mieux les prédictions des modèles primaires.)
- Support Vector Machines
  - Classification par hyper-plan séparateur ; classifieur de marge maximale ; données non linéairement séparable et méthodes à noyau ; SVM multi-classe ; liens avec d'autres modèles (logistique, réseaux de neurones) ; descente de gradient

Compétences

- Identifier comment résoudre une tâche par apprentissage automatique
- Choisir un modèle a priori adapté à une tâche
- Utiliser un modèle de l'état de l'art (SVM, réseau de neurones, forêt, …)
- Comparer empiriquement différents modèles pour une tâche donnée

Pré-requis

R, Python, algèbre linéaire, optimisation de fonctions

Contrôle des connaissances

Des TP notés + un examen final

Références bibliographiques

- Endrew Ng. Machine Learning Yearning. Disponible gratuitement au lien https://www.deeplearning.ai/machine-learning-yearning/.
- Rémi Gilleron. Apprentissage machine - Clé de l'intelligence artificielle - Une introduction pour non-spécialistes. Ellipses.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. 2016

Langue d'enseignement

Français

UE1 – Machine Learning

Machine Learning – Réseaux de neurones avancés

Machine Learning

Cours : 3h • Atelier : 9h

Enseignant : Romaric GAUDEL (Ensai)
Correspondant : Romaric GAUDEL (Ensai))

Objectif pédagogique

Ce cours s'intéresse à des réseaux de neurones aux architectures plus complexes.

Contenu de la matière

Connaissances

- Réseaux de neurones pour séries temporelles
- Modèles Génératifs (auto-encodeurs (variationnels), réseaux antagonistes génératifs)

Compétences

- Mise en application des modèles étudiés

Pré-requis

Python, principes de l'apprentissage automatique, réseaux de neurones

Contrôle des connaissances

TP notés

Références bibliographiques

- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. 2016

Langue d'enseignement

Français

UE1 – Machine Learning

Machine Learning – Systèmes de recommandation

Machine Learning

Cours : 6h • Atelier : 6h

Enseignant : Romaric GAUDEL (Ensai)
Correspondant : Romaric GAUDEL (Ensai)

Objectif pédagogique

Les systèmes de recommandation choisissent les options à présenter à des utilisateurs parmi un grand nombre de possibilités. Ils permettent par exemple de recommander la prochaine vidéo à regarder, le prochain morceau à écouter, les photos à montrer… Le cours présentera les modèles utilisés pour construire de tels systèmes.

Contenu de la matière

- Objectif des systèmes de recommandation
- État de l'art des systèmes de recommandation (plus proches voisins, filtrage collaboratif…)
- Évaluation des systèmes de recommandation
- Problèmes rencontrés par les systèmes de recommandation et solutions afférentes (démarrage à froid, compromis exploration-exploitation...)

Compétences

- Mise en application des modèles étudiés
- Évaluation d'un système de recommandation

Pré-requis

Python, principes de l'apprentissage automatique

Contrôle des connaissances

TP notés

Références bibliographiques

- Statistical Methods for Recommender Systems. Deepak K. Agarwal, Bee-Chung Chen. 2016.
- Recommender Systems: The Textbook. Charu C. Aggarwal. Springer, 2016.
- Bandit algorithms for Website optimization. John Myles White. O'Reilly Media.
- Blog et tutoriels de Sebastien Bubeck's : https://blogs.princeton.edu/imabandit/
- Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Sébastien Bubeck, Nicolò Cesa-Bianchi. https://arxiv.org/abs/1204.5721
- Bandit Algorithms. Tor Lattimore and Csaba Szepesvári. Cambridge University Press. https://tor-lattimore.com/downloads/book/book.pdf

Langue d'enseignement

Français

UE1 – Machine Learning

Régression pénalisée et sélection de modèles

Penalized problems and model selection

Cours : 9h • Atelier : 6h

Enseignants : Cédric HERZET (INRIA) & Clément ELVIRA (INRIA)
Correspondant : Arthur KATOSSKY (Ensai)

Objectif pédagogique :

De nombreuses tâches d'apprentissage et de traitement du signal visent à retrouver un ensemble de grandeurs inconnues (état d'un système, modèle génératif, etc) à partir de données.

Malheureusement, dans de nombreuses situations, les données disponibles s'avèrent insuffisantes pour lever l'ambiguïté sur les quantités à inférer ou les estimer avec une précision suffisante.
Une manière classique de contourner ce problème consiste à introduire une information « a priori » sur la solution recherchée.
Plus particulièrement, dans ce cours nous montrons comment lever l'ambiguïté inhérente à certains problèmes en « pénalisant » les solutions ne présentant pas certaines caractéristiques d'intérêt.
L'objectif de ce cours est d'identifier et manipuler les méthodes de pénalisation les plus courantes.

Contenu de la matière

- Identifier la pénalisation la plus adaptée à une tâche
- Résoudre un problème d'optimisation comportant un terme de régularisation
- Régler les paramètres du modèle

Pré-requis

- Algèbre linéaire
- Optimisation convexe
- Programmation en Python

Contrôle des connaissances

TP notés + examen final

Références bibliographiques

- C. Bishop. Pattern recognition and machine learning. Springer-Verlag New York, 2006.
- S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhau¨ser, 2013.
- D. P. Bertsekas. Nonlinear Programming. Athena Scientific, USA, 2003.

Langue d'enseignement

Français

UE1 – Machine Learning

Apprentissage statistique à grande échelle

Large-scale Machine-Learning

Cours : 9h • Atelier : 9h

Enseignant : Arthur KATOSSKY (Ensai) et Rémi PEPIN (Ensai)
Correspondant : Arthur KATOSSKY (Ensai)

Objectif de la matière

Au cours de la dernière décennie, nous avons assisté à l'émergence d'applications numériques nécessitant de faire face à de gigantesques quantités de données, générées de plus en plus rapidement. Ces applications (surveillance de réseaux, biologie et médecine, applications financières, réseaux sociaux, etc.) nécessitent un besoin grandissant de techniques capables d'analyser et de traiter ces grandes masses d'information, avec précision et efficacité. La statistique rejoint ici les sciences du numérique, et plus précisément l'informatique répartie, pour proposer de nouvelles approches, relatives au Big Data. Les techniques et les modèles doivent prendre en compte le volume pléthorique de ces données, mais également leur génération rapide en continu (vélocité) ainsi que la diversité de leur format (variété) et la qualité de l'information (véracité), appelés communément les 4V du Big Data.

Contenu de la matière

- Les différents « v »
- Principes, avantages et inconvénients d'un système réparti
- connaître les stratégies de tolérance aux fautes (duplication des données, exécution avec erreurs)

Compétences

- Identifier l'architecture adaptée à une tâche (exécution séquentielle et/ou parallèle, exécution en mémoire et/ou en flux, exécution locale et/ou distante).
- Lancer des calculs sur une architecture Big Data (notamment, appliquer les paradigme Map-Reduce).
- exécuter des calculs volumineux – et en particulier des calculs statistiques – sur des prestataires de calcul (IaaS ou PaaS comme Amazon Web Service, Google Cloud Platform ou autre)

Pré-requis

Algorithmique.

Contrôle des connaissances

À déterminer.

Références bibliographiques

- Analyses des Big Data : quels usages, quels défis ? Note d'analyse du Comissariat général à la stratégie et la prospective
- Pirmin Lemberger, Marc Batty, Médéric Morel, Jean-Luc Raffaëlli. Big Data et machine learning - Manuel du data scientist, Dunod, 2015.
- Rudi Bruchez. Les bases de données NoSQL et le BigData : Comprendre et mettre en œuvre, Eyrolles (2015)

Langue d'enseignement

Français.

UE1 – Machine Learning

Webmining et traitement du langage

Webmining et NLP

Cours : 9h • Atelier : 12h

Enseignant : Arthur KATOSSKY (Ensai)
Correspondant : Arthur KATOSSKY (Ensai)

Objectif pédagogique :

Le cours de *webmining & natural language processing* répond à plusieurs objectif:
- pratiquer la collecte de données, l'extraction d'information et l'appariement de sources
- équiper les élèves avec des outils théoriques pour l'étude des données textuelles
- faire comprendre les grandes approches qui structurent le foisonnement de modèles de la langue
- présenter des exemples concrets d'applications dans les différentes domaines d'application des élèves
- donner la capacité de réaliser des tâches classiques en étude de texte: classification, analyse de sentiment, détection d'entités, etc.

Contenu de la matière

- Introduction au traitement automatique du langage (*natural language processing*)
- Grandes catégories de modèles : *bag-of-words* et *tf-ifd* ; réseaux de neurones (LSTM, GRU, etc.) ; plongements de mots (word2vec, GLoVe, fasttext, Elmo, BERT, etc.) ; modèles probabilistes (HMM, CRF, LDA, etc.)
- Applications : classification, analyse de sentiment, détection d'entités, etc.
- Traitement de donnés textuelles et extraction d'information
- Collecte de données sur le web et utilisation d'une API

Pré-requis

Apprentissage statistique (réseaux de neurones) ; apprentissage statistique à grande échelle ; statistique bayésienne ; chaînes de Markov

Contrôle des connaissances

Projet

Références bibliographiques

Communiquée ultérieurement

Langue d'enseignement

Français.

Descriptifs des enseignements de la filière

UE2 – Développement d'application

Génie Logiciel

Software engineering

Cours : 39h • Atelier : 39h

Enseignant : Mathieu ACHER, Johann BOURCIER , Olivier BARAIS, et Benoit COMBEMALE (Université Rennes 1), Mohamed GRAIET (Ensai)

Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

L'objectif de ce cours est d'introduire les moyens de concevoir des applications informatiques de qualité (répondant aux besoins, évolutives et faciles à maintenir).
Il s'agit de présenter l'ingénierie dirigée par les modèles en positionnant la conception dans les cycles de développement, et en mettant l'accent sur les enjeux et les pièges à éviter.
Le cours présente une introduction aux modèles de conception classiques, base du génie logiciel autour des technologies objet, en proposant des applications pratiques au cours de travaux pratiques et en étudiant des patrons de conception développés en Java. Cet enseignement vise également à apprendre à développer et déployer un site Web dynamique en Java. Il permet de se familiariser avec les architectures n-tiers et les serveurs d'applications et de bien maîtriser les principaux outils et langages avancés de développement des applications Web/JavaEE.

Contenu de la matière

I. **Le génie logiciel**
- Pourquoi ?
- Comment ?

Introduction au génie logiciel et bonne pratiques de conception.
Architecture logicielle et modèle en couche, Exemple sur GWT.
Principaux patrons de conception, principe et mise en œuvre en Java.
Le test logiciel et l'ingénierie des langages.
L'ingénierie dirigée par les modèles.

II. **Programmation Client Serveur (JavaEE)**
1. Architectures distribuées et plate-forme JavaEE
   i. Les Technologies JavaEE et Spring
   ii. Architecture : composants, services et communications
   iii. Les problématiques des applications serveurs
2. API et frameworks JavaEE / Spring
3. La persistance avec JPA
   i. Problématique du "mapping" objet-relationnel
   ii. Les outils de mapping : JPA, Hibernate, Toplink
   iii. Le mapping
   iv. L'entity-manager
   v. Le langage de requetage
4. Les services web, le cloud.

Pré-requis

Notation UML, connaissance du langage JAVA.

Contrôle des connaissances

Un TP noté sera rendu.

Références bibliographiques

- I. SOMMERVILLE, *Le Génie logiciel*, Addison Wesley-France, 1988
- B. BEIZER, *Software Testing Techniques*, Second Edition, Van Norstrand, 1990
- B.W. BOEHM, *Software Engineering Economics,* Prentice-Hall, 1981
- E. GAMMA, R. HELM, R. JOHNSON, J. VLISSIDES, *Design patterns, catalogue de modéles de conception réutilisables*, Vuibert, 2007

Langue d'enseignement

Français

UE2 – Développement d'application

Indexation Web

Web datamining

Cours : 9h • Atelier : 6h

Enseignant : Nawfal TACHFINE (aramisauto)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

A l'issue de ce cours, les élèves devront savoir collecter des informations issues du web, connaître la notion d'Information Retrieval, savoir constituer des corpus, et les organiser à des fins d'analyse exploratoires. Ils devront maîtriser également l'algorithme qui permet de hiérarchiser les pages web (pagerank) et les techniques de classification de documents textuels.
Par ailleurs, ils devront avoir acquis les notions d'opinion mining (classification de textes, analyses de sentiments, évaluation de modèles).
Toutes les applications seront traitées en R.

Contenu de la matière

Partie 1 – Information Retrieval : Preprocessing, Extraction and PageRank

Mots clés : *twitter, R, pagerank, corpus, term-document matrix, Information retrieval, tf-idf, stemming, Regex, kmeans*

Partie théorique (3h)

- Information Retrieval
    - Concepts & Définitions
    - Term Document Matrix
    - Tf-idf, Cosine Index, jaccard Index
    - Stemming

- Web Search : Google
    - Google et le Page Rank
    - Pages Jaunes (Notion de tri alpha)
    - Notion de graphes et de vecteurs propres

Partie pratique (9h)

- TP1 : Introduction à R pour le Web Mining (3h)
    - Installation de librairies de textmining disponible dans R
    - Collecter les informations issues du WEB : Twitter, Wikipedia
    - Pre-processing : Stemmatisation, Lemmatisation,
    - Parsing HTML, XML,
    - Tokenization
    - Introduction à la term-document matrix
- TP2 : Similarité de documents (Applications aux recherches utilisateurs sur le site pagesjaunes.fr (3h)
    - Indices de similarité : Tf, tf-idf Jaccard, Cosine
    - Distance de Damerau, Distance de jaro
    - Liens entre les recherches, Notion de graphe de recherche
- TP3 : Ordonnancement des résultats d'une recherche (3h)
    - PageRank
    - Détecter les mots clés
    - Intro à la classification des docs sur mots clés

Partie 2 – Opinion Mining : Textmining, analyse de sentiments, classification et évaluation des modèles.

Mots clés : *Facebook, R, opinion mining, corpus, sentiment analysis, annotation syntaxique.*

Partie théorique (4h)

- Introduction
  - Quelles applications dans quels domaines d'activités
- État de l'art (opinion mining, sentiment analysis, affective computing)
  - Quels descripteurs pour quels types de données ?
    - Textuelles
    - Audio
    - Images
  - Sélection automatique de descripteurs (réduction de l'espace de recherche)
  - Quels algorithmes de classification dans quels cas ?
- Constitution du corpus
  - Réflexions générales sur la qualité des données et son impact
  - Annotation manuelle et automatique (schéma d'annotation, calcul d'un score d'agrément inter-annotateur,)
  - Répartition des données dans les classes
- Pre-processing (texte)
  - Quelle granularité pour mes données (mot, phrases, paragraphes)
  - Annotation syntaxique et sémantique (exemples de POS, WordNet-Affect, etc)
- Évaluation
  - Quelles mesures utiliser pour mesurer la qualité d'un modèle (rappel, précision, f-score, ROC, indices de confiance a 0.95)
- Les produits du marché (exemples)
  - Produit de la société TEMIS (cartouche sentiments)
  - Produit de la société Sinequa

Partie pratique (8h)

- TP1 : classification de la valence d'un texte littéraire (critiques de cinéma)
- TP2 : classification de la valence de textes issus de réseaux sociaux (twitter, facebook)
- TP3 : Fusion de modèles (à partir des modèles crées dans le TP2)
- ***TP4 (optionnel)*** *: Constructions de modèles à partir d'indices multimodaux (texte + audio)*

Pré-requis

SQL.

Contrôle des connaissances

Projet par groupe d'élèves.

Références bibliographiques

Les * indiquent les lectures fortement conseillées.

- Web DataMining, Exploring Hyperlinks, Contents, and Usage Data, Bing Liu, Springer (Chapitre 6 à 13) (*)
- Information Retrieval, http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf (chapitres 1-3) (*)
- **package tm in *R***, http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf (*)
- Infrastructure of Texmining with R, http://www.jstatsoft.org/v25/i05/paper
- Webmining pluging in R, http://cran.r-project.org/web/packages/tm.plugin.webmining/vignettes/ShortIntro.pdf
- PageRank, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
- Introduction to PageRank, http://www.stat.berkeley.edu/~vigre/undergrad/reports/christensonNathan.pdf (*)
- Mining the social web, https://github.com/ptwobrussell/Mining-the-Social-Web

- Pang B. and Lee L. (2008). "Opinion mining and sentiment analysis." Foundations and Trends in Information Retrieval **2**(1-2).
- Dini L. and Mazinni G. (2002). Opinion classification through information extraction. CELI. Turin, Italy
- Cornuéjols A., Miclet L. and Kodratoff Y. (2002). Apprentissage artificiel : Concepts et algorithmes
- Ilieva L. (2004). Combining Pattern Classifiers : Methods and Algorithms (chapitre 1 “Fundamentals of Pattern Recognition”, chapitre 4 “Fusion of Label Outputs”) (*)

Langue d’enseignement

Français.

UE2 – Développement d'application

Projet WEB et Applications WEB

Web applications

Cours : 12h • Atelier : 15h • Projet : 9h

Enseignants : Olivier BARAIS (U. Rennes I) & Olivier CHANTREL (Orange)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

L'objectif de cet enseignement est d'effectuer un projet de bout en bout. Ce projet commence par une modélisation utilisant les méthodes et techniques vues en génie logiciel et se termine par une implémentation en JavaEE.
Ce cours vise à donner aux étudiants une vision détaillée du web d'aujourd'hui en présentant les technologies historiques du web (html, xml) et les technologies plus récentes du web 2.0 (css, javascript, HTML5, Ajax, Php, MySql).

Contenu de la matière

- Encadrement en début de projet afin de préciser les besoins et pour déterminer l'architecture générale du programme.
- Encadrement distant sur des questions techniques ponctuelles
- Encadrement technique lors de la phase d'implémentation
- Soutenance des projets

- L'historique du web / XML et ses applications (html, xml, dtd, web services etc.)
- Le web actuel (css, javascript, HTML5, Ajax, MySql, Php)

Contrôle des connaissances

Les étudiants sont évalués sur la base du rapport d'étude et d'une soutenance devant un jury, incluant une démonstration de leur réalisation.

Références bibliographiques

- L. Roland, « Structurez vos données avec XML », 2014
- L Van Lancker, « AJAX - Développez pour le Web 2.0 - Entrez dans le code : JavaScript, XML, DOM, XMLHttpRequest2... (2ième édition) », 2015
- C Pierre de Geyer & E Daspect, « PHP5 avancé », 2012

Langue d'enseignement

Anglais

UE3 – Big Data

Technologies Sémantiques

Semantic Technologies

Cours : 6h • Atelier : 9h

Enseignants : Sébastien FERRE (IRISA)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique:
Le cours vise à donner aux étudiants une vision détaillée de la prochaine génération du web - le web sémantique -, qui introduit le sens de l'information dans les échanges de données pour permettre aux machines de traiter automatiquement l'information disponible sur le web. Le cours présente les standards du web sémantique et propose aux étudiants de manipuler des outils implémentant ces standards pour répondre à un cas d'usage concret.

Contenu de la matière:
* Introduction au Web sémantique et au modèle de description RDF
* Introduction à RDFS/OWL et au langage SPARQL
* Le Web sémantique en pratique

Pré-requis:
* Programmation Java
* Connaissance des technologies du web: HTTP, HTML, XML

Contrôle des connaissances:
TP noté

Références bibliographiques:
* Grigoris Antoniou and Frank van Harmelen, A Semantic Web Primer, 2nd Edition (Cooperative Information Systems), 2008

Langue d'enseignement:
Français

UE3 – Big Data

Technologies NoSQL

NoSQL Technolgies

Cours : 12h • Atelier : 3h

Enseignant : David GROSS-AMBLARD (IRISA)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

Ce cours vis à présenter les différentes approches présententes dans le contexte des bases de données NoSQL. Ces bases de données se distinguent des approches classiques relationnelles. Ces approches abandonnent la représentation matricielle de l'information ainsi que le langagage SQL au profit d'une plus grande simplicité, d'une meilleure performance et d'une meilleure scalabilité.

Contenu de la matière

Pré-requis

Bases de données relationnelles

Contrôle des connaissances

Références bibliographiques

Langue d'enseignement

Français

UE – Big Data

Publication de données respectueuse de la vie privée

Privacy-preserving data publishing

Cours : 15h • Atelier : 6h

Enseignant : Tristan ALLARD (Univ. Rennes 1)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif de la matière

« Les données personnelles sont le nouveau pétrole d'Internet et la nouvelle monnaie du monde numérique » a déclaré M. Kouneva, commissaire européen à la protection des consommateurs en mars 2009. La valeur de l'analyse massive des données personnelles pour les industriels, les scientifiques et la société en général est largement reconnue aujourd'hui. Cependant, leur caractère personnel et potentiellement sensible est un obstacle majeur à leur partage à grande échelle. L'objectif des modèles et algorithmes de publication de données respectueuse de la vie privée est précisément d'offrir des garanties fortes de respect de la vie privée tout en autorisant un partage de qualité à des fins d'analyse. La tâche est loin d'être triviale comme l'ont démontré plusieurs scandales de ré-identification. L'objectif de ce cours est de présenter aux étudiants les principaux paradigmes et techniques de publication de données respectueuse de la vie privée.
L'accent sera particulièrement mis sur un modèle proéminent aujourd'hui : la differential privacy.

Contenu de la matière :

- Introduction : motivation, défis, survol
- Paradigmes : non-informatif, differential privacy
- Publication interactive: modèles type differential privacy, mécanismes principaux de perturbation intéractive (e.g., Laplace)
- *Perturbation locale : le mécanismes des réponses randomisés pour satisfaire la differential privacy
- Publication centralisée : mécanismes de génération de données synthétiques satisfaisant la differential privacy, survol des modèles basés sur le partitionnement (e.g., k-anonymat, l-diversité) et des mécanismes principaux pour les satisfaire (e.g., algorithme de Mondrian)
- Conclusion : les pratiques « dans le monde réel », questions ouvertes

Pré-requis :

- Connaissances de base en gestion de données, en algorithmique, et en probabilités et statistiques.
- Compétences de base dans un langage de programmation parmi Java, Python, ou R.

Contrôle des connaissances :
Contrôle continu et examen final.

Références bibliographiques :

- B.-C. Chen, D. Kifer, K. LeFevre, et A. Machanavajjhala, Privacy-Preserving Data Publishing, Found. Trends databases, vol. 2, no 1-2, p. 1-167, 2009.
- C. Dwork et A. Roth, The Algorithmic Foundations of Differential Privacy, Found. Trends Theor. Comput. Sci., vol. 9, no 3-4, p. 211-407, 2014.
- B. C. M. Fung, K. Wang, R. Chen, et P. S. Yu, Privacy-preserving data publishing : A survey of recent developments, ACM Comput. Surv., vol. 42, no 4, p. 14:1-14:53, 2010.

Langue d'enseignement:
Français

UE4 – Systèmes et Réseaux

Réseaux et systèmes d'exploitation

Computer Networks and Operating System

Cours : 15h • Atelier : 6h

Enseignant : Jean-Baptiste LOISEL (Orange Consulting)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données » et du Master Big Data

Course Objectives

This course aims to provide students with an understanding of the core principles of technologies constituting the foundation of the IT world: operating systems and computer networks.
In the first part, we will study the way an operating system organizes and facilitates the interaction of its key resources such as processor, memory, and file system in a multi-tasking and multi-user context.
The second part will focus on networks and will address various topics, such as network topology and technologies, Ethernet, ADSL, LAN, WAN, VLAN, Internet, Wifi and secure Wifi, TCP/IP layers, major protocols (DNS, SMTP…), network devices, architecture designs (dimensioning, redundancy, segmentation, DMZ…).
Implications for the security of the Information System will also be touched when addressing these topics, in order to raise awareness about inherent security risks and relevant countermeasures.

Course description

Operating Systems
1. Operation Systems overview
2. Operation Systems overview
3. Processes
4. Inter-process communication
5. Memory management
6. Processes scheduling
7. File systems
8. Disk management systems (RAID)
9. Virtualization

Computer Networks
1. Introduction
2. Host-network layer
3. Internet layer
4. Transport layer
5. Application layer
6. Architecture review

Practicals will supplement the course.

Course evaluation

Written exam

Bibliography

- Modern Operating Systems. Andrew Tanenbaum. Pearson Education. 4th edition (2014). ISBN-13: 978-0133591620 ISBN-10: 013359162X
- Computer networks. Andrew Tanenbaum & David Wetherall. Pearson. 5th edition (2010). ISBN-13: 978-0132126953 ISBN-10: 0132126958

Langue d'enseignement

Anglais

UE4 – Systèmes et Réseaux

Initiation à Unix

Networks and Systems

Cours : 9h • Atelier : 6h

Enseignant : François Xavier BRU (Orange Consulting)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

Il s'agit d'un atelier intense pendant lequel les étudiants vont installer une version récente de Linux et apprendre à manipuler ce système d'exploitation afin de l'utiliser tout au long de l'année.
Linux est en particulier central pour utiliser et développer les technologies Big Data.

Contenu de la matière

1. Présentation d'Unix
2. Installation d'une version
3. Découverte pratique d'unix
4. Installation de logiciels

Pré-requis

Contrôle des connaissances

Le contrôle des connaissances s'effectue sur l'ensemble des matières de l'UE.

Références bibliographiques

C. PELISSIER, Unix, Editions Hermès

Langue d'enseignement

Français

UE4 – Systèmes et Réseaux

Systèmes Répartis

Networks

Cours : 15h • Atelier : 6h

Enseignant : David FREY & George GIAKKOUPIS (Inria Rennes)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

Cet enseignement vise à donner aux étudiants les connaissances de base sur les architectures distribuées, réparties sur différents sites partout dans le monde. Les trois architectures réparties à grande échelles les plus courantes seront présentées : grilles, systèmes peer-to-peer, et cloud. Les hypothèses, concepts and algorithmes seront détaillés pour chacune d'entres elles. L'objectif est d'avoir une connaissance des systèmes répartis disponibles actuellement et de pointer les directions futures de ces architectures.

Contenu de la matière

1. Introduction aux architectures distribuées
2. Les concepts fondateurs (synchronisation, exclusion mutuelle, etc.)
3. Les approches centralisées et semi-centralisées (cloud, grilles, etc.)
4. Les approches décentralisées (systèmes P2P - structurés, non-structurés et hybrides)
5. Application aux systèmes de partage de fichiers et aux protocoles épidémiques

Pré-requis

Algorithmique, Programmation orientée Objet

Contrôle des connaissances

Un projet de mise en œuvre de système décentralisé, type épidémique

Références bibliographiques

- Andrew S. Tanenbaum et Maarten Van Steen. Distributed Systems: Principles and Paradigms. Pearson New International Edition (2013)
- Kenneth Birman . Guide to Reliable Distributed Systems. Springer Verlag (2012)
- Fabrice Le Fessant et Jean-Marie Thomas. Le peer-to-peer : Comprendre et utiliser. Eyrolles (2011)
- Andrew S. Tanenbaum. Systèmes d'exploitation : Systèmes centralisés, systèmes distribués. Dunod (1999)

Langue d'enseignement

Français

UE4 – Systèmes et Réseaux

Sécurité des données

Data Security

Cours : 9h • Atelier : 6h

Enseignant : Franck LANDELLE (DGA MI)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

La sécurité informatique fait actuellement l'objet d'une actualité particulièrement dynamique : attaques spectaculaires (virus, intrusion, ...), commerce électronique, évolutions de législations…
L'objet de ce cours est de présenter les grands principes de la sécurité informatique et les techniques de protection des données.
L'usage de la cryptographie est l'un des outils de protection contre la divulgation, la modification ou l'accès illégitime à des données ou moyens.
Les techniques cryptographiques qui permettent d'assurer les services de confidentialité, d'intégrité, de signature ou d'authentication.
Finalement, des systèmes utilisant ces techniques seront schématiquement décrits.

Contenu de la matière

1. Introduction à la sécurité
   1.1. Besoins
   1.2. Menaces
2. Cryptographie
   2.1. Définitions générales
   2.2. Cryptographies à clés secrètes
   2.3. Cryptographies à clés publiques
   2.4. Protocoles cryptographiques
3. Systèmes utilisateurs
   3.1. Applications Web
   3.2. Carte bancaire
3.3. Application réseaux

Pré-requis

Contrôle des connaissances

Examen écrit

Références bibliographiques

- Schneier, Cryptographie appliquée, Thomson Publishing, 1997
- Stinson, Cryptographie : Théorie et pratique, Vuibert 2003
- Menezes, Van Oorschot, Vanstone, Handbook of applied Cryptography, CRC Press, 1997 (version actualisée en ligne)
- Vergnaud, Exercice et problèmes de la cryptographie, Dunod, 2012.
- Singh, Histoire des codes secrets, JC Lattes, 1999

Langue d'enseignement

Français

UE4 – Systèmes et Réseaux

Grandes masses de données sur Cloud

Big Data and Cloud Computing,

Cours : 12h • Atelier : 12h

Enseignant : Gabriel ANTONIU (Inria Rennes)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif de la matière

Cet enseignement vise à donner aux étudiants les connaissances de base sur les architectures distribuées spécialisées dans le traitement du Big Data. Les hypothèses, concepts and algorithmes seront détaillés pour chacune d'entres elles. L'objectif est d'avoir une connaissance des systèmes sur Cloud disponibles actuellement et de pointer les directions futures de ces architectures. L'objectif final est la mise en œuvre avec Hadoop et une base de données Big Data spécifique.

Contenu de la matière

- Introduction aux infrastructures distribuées : clusters, supercalculateurs, grilles, clouds
- Introduction au cloud computing
- Big Data: introduction, défis, enjeux
- Explicit Data management
- Transparent Data Management: NFS, Gfarm, Google File System:
- Introduction à MapReduce et Hadoop:
- Atelier pratique sur Hadoop
- Avenir de MapReduce: défis
- Approches post-MapReduce: Shark/Spark (Berkeley)
- Approches post-MapReduce: Stratosphere

Pré-requis

Système répartis, interrogation (SQL) de bases de données relationnelles.

Contrôle des connaissances

Examen écrit.

Références bibliographiques

- G. PLOUIN, Cloud computing et SaaS, Editions Dunod
- Le livre blanc du Cloud, du SaaS et des Managed Services pour les partenaires IT et télécoms. Edition 2013
- R. HENNION, H. TOURNIER, E. BOURGEOIS, Cloud computing : Décider - Concevoir - Piloter - Améliorer, Editions Eyrolles, 2012

Langue d'enseignement

Français

UE – Projet de fin d'études

Projet méthodologique – Veille sur les médias

Methodological project - Technology watch through medias

Atelier : 3h • Projet : 9h

Enseignant : Mohamed GRAIET (Ensai)
Correspondant : Mohamed GRAIET

Enseignement destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

L'évolution du monde informatique est constante, tant au niveau des techniques de logiciel que des ressources matérielles. Il se développe sans cesse un « jargon » informatique, de nouveaux sigles apparaissent régulièrement dans la presse spécialisée, les sociétés travaillant dans le domaine de l'informatique sont l'objet de rachats, regroupements stratégiques… L'objectif de ce projet est d'habituer les étudiants à se maintenir en état de « veille» à travers la « grande » presse informatique pour être au courant des technologies qui arrivent, des nouvelles évolutions du marché des SSII par exemple, et pour pouvoir rapidement s'adapter à de nouveaux environnements.

Contenu de la matière

Chacun des étudiants de la filière doit réaliser une étude sur un sujet qu'il choisit en accord avec l'encadrant :
- étudier comment une nouvelle technologie est vue à travers plusieurs medias
- étudier sur plusieurs semaines comment évolue l'information concernant une technologie dans un media donné
- approfondir un sujet exposé dans un media.

Pré-requis

Contrôle des connaissances

Ce projet donnera lieu à une soutenance et des discussions avec l'ensemble des étudiants.

Références bibliographiques

Langue d'enseignement

Français

UE – Projet de fin d'études

Projet de fin d'études

End of study project

Atelie : 9h • Projet : 27h

Enseignant : Divers intervenants industriels
Correspondant : Arthur KATOSSKY (Ensai)

Enseignement destiné à l'ensemble des élèves des six filières

Objectif pédagogique

Le projet de fin d'études consiste en la production d'une étude statistique de niveau professionnel dans le monde de l'entreprise ou de la recherche, parmi un catalogue de sujet mis à disposition des élèves. Ses objectifs sont multiples:
- mise en situation professionnelle
- capacité à définir une stratégie d'étude en réponse à une demande client
- mobilisation des compétences techniques (statistiques, économiques, informatiques)
- compromis entre rigueur scientifique et contraintes pratiques (limitations financières, logicielles, cognitives, temporelles…)

travail de groupe
- gestion d'un projet sur le temps long
- communication (écrite, orale) sur des sujets techniques

Contenu de la matière

Travail autonome en groupe suivi par un professionnel de l'entreprise ou de la recherche (env. 5 séances)

Pré-requis

Références bibliographiques

Selon les projets

Contrôle des connaissances

Évaluation: projet avec soutenance

Langue d'enseignement

Français

UE – Projet de fin d'études

Data challenge

Data Challenge

Atelier : 12h

Enseignant : Divers intervenants industriels
Correspondant : Salima EL KOLEI

Enseignement destiné à l'ensemble des élèves des six filières

Objectif du cours électif

Le data challenge permet de rassembler sur une période très courte différentes équipes de profils variés afin de collaborer sur un projet. Cette expérience se rapproche des conditions réelles dans laquelle évoluent les datascientists au sein des entreprises.
Il permet, à partir des mécanismes du jeu, de dynamiser et d'articuler la pédagogie autour d'un besoin concret d'entreprise et d'un événement qui s'achève par une évaluation objective. De nombreux challenges sont proposés autour de la Data ou présentant des problématiques Data importantes.
L'objectif de ce cours est de valoriser et d'évaluer les compétences transversales acquises dans ce contexte opérationnel.

Contenu de la matière

Les élèves devront participer au data challenge proposé à l'Ensai ouvert également aux élèves de deuxième année.

Compétences acquises

-Comprendre les problèmes à résoudre.
-Travailler en mode projet avec des contraintes.
-S'intégrer et s'adapter dans un contexte pluridisciplinaire. Selon les challenges, les compétences seront mobilisées à géométrie variable.
-S'adapter à la réalité de la Data d'entreprise (données non structurées, manquantes, volumétrie…)
-Communication orale des résultats (pitch…)

Prérequis

-Compétences en statistiques et informatiques de 1A, 2A et 3A.
-Compétences transversales mobilisées dans les projets 1A, 2A et 3A.

Séminaires professionnels

Séminaires professionnels

Seminars

Cours : 30h

Enseignants : Divers intervenants
Correspondant : Mohamed GRAIET

Séminaires destiné aux élèves de la filière « Statistique et Ingénierie des Données »

Objectif pédagogique

Découvrir de façon ludique et sans contrôle de connaissance des technologies importantes qui ne peuvent pas être intégrées dans le cursus classique de la formation.
Ces technologies sont abordées par l'exemple, directement par la pratique des entreprises ou des laboratoires de recherche
Ces séminaires thématiques seront complétés, au cours de l'année, par des séminaires de présentation des problématiques actuelles de différentes enteprises.

Contenu de la matière

Spark
ElasticStack
TensorFlow
Hive

Pré-requis

Contrôle des connaissances

Aucun.

Références bibliographiques

Langue d'enseignement

Français