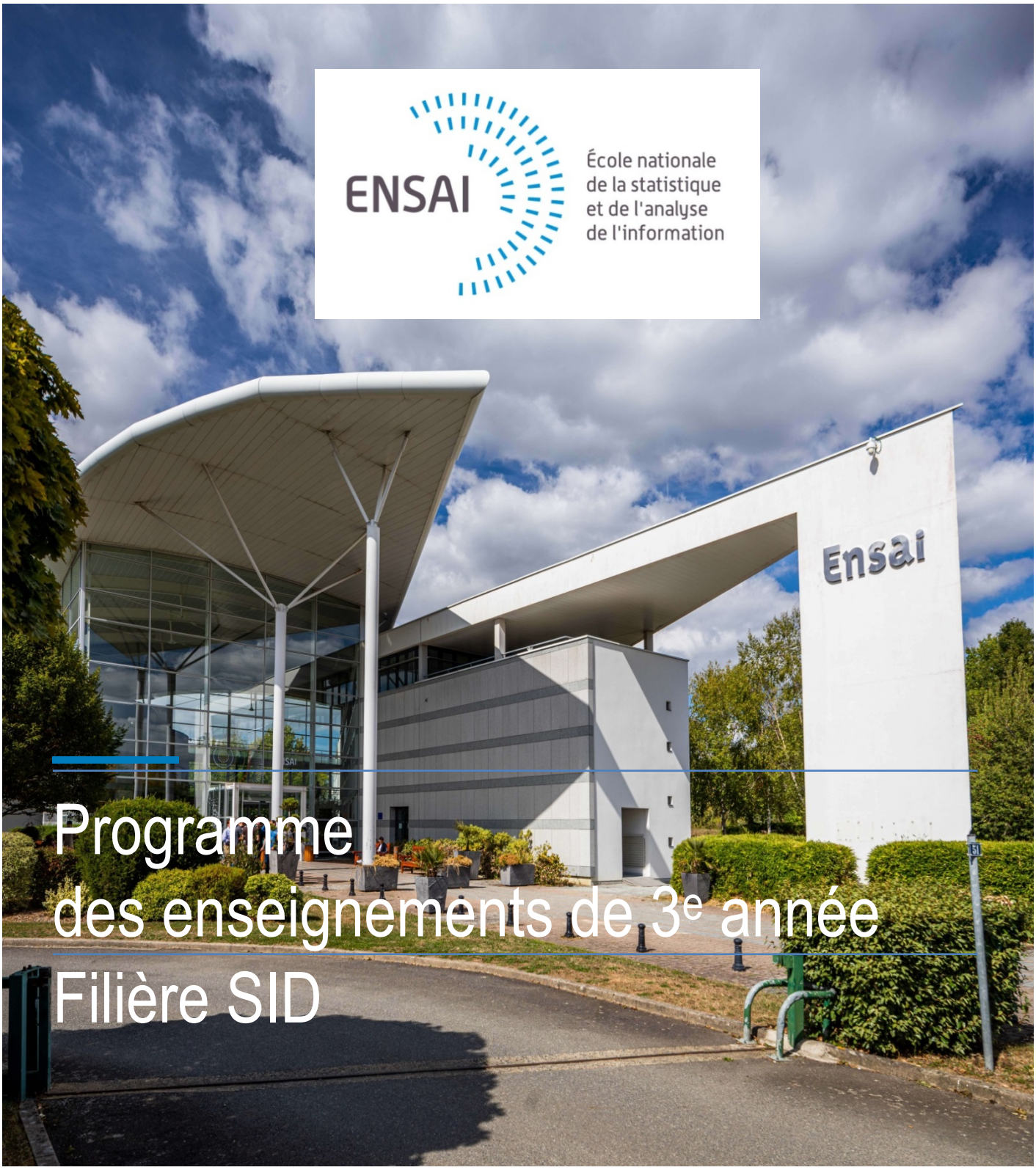

Programme des enseignements de 3e année Filière SID

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

Campus de Ker Lann, 51 rue Blaise
Tél : 33 (0)2 99 05 32 32 / scolarite@

www.ensai.fr

FILIÈRE STATISTIQUE ET INGÉNIERIE DES DONNÉES

ANNÉE SCOLAIRE 2020/2021

Data Science and Big Data

2020/2021 ACADEMIC YEAR

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Présentation de la filière

La formation d'ingénieur de l'Ensai inclut 6 filières de spécialisation. Toutes ces filières forment aux métiers de la Data Science, avec une maîtrise des outils permettant l'extraction, l'analyse et la fouille de données et une capacité à choisir les modalités de traitements des données massives (Big Data) et des techniques d'apprentissage automatique (machine learning). Selon les spécialisations, ces compétences sont spécifiques à un domaine ou transversales. Ces compétences sont renforcées pour la filière Statistique et Ingénierie des Données (SID), qui s'est concentré sur ces thématiques depuis plusieurs années. L'ensemble des filières continue à former aux compétences transversales (soft skills) et à la valorisation des travaux menés dans un contexte professionnel et international. Lors des cours et du projet méthodologique en anglais, les élèves travaillent toutes les compétences linguistiques et communicationnelles et approfondissent leurs connaissances liées au monde de l'entreprise et de la recherche. La séquence de Tronc Commun mêlant enseignements scientifiques, projets et anglais conclut la formation à l'autonomie et la capacité à mettre en œuvre des analyses de données en situation complexe. Un stage de fin d'études est à réaliser à l'issue de la scolarité, qui permet de mettre en œuvre dans un cadre professionnel une démarche scientifique autour d'une problématique en lien avec les enseignements de la filière.

Les enseignements spécifiques de la filière se divisent en deux, les enseignements informatiques et les enseignements liés à l'ingénierie des données.

Le développement de grandes applications informatiques nécessite d'utiliser des méthodes d'aide à la conception. Elles peuvent être purement techniques, comme la maîtrise de patron de conception ou l'architecture multi-tiers ou organisationnel comme les méthodes de gestions de projet, aussi bien pour la spécification d'un besoin que le suivi. L'unité s'enseignement (UE) **développement d'application** apportera aux élèves ces compétences essentielles à de futurs chefs de projet informatique. L'UE de **système et réseaux** quant à elle se concentrera sur la conception, la mise en place et l'administration des systèmes complexes pour traiter des données.

En parallèle, 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. L'UE de ***Big Data*** va rentrez va présenter les 3 dernières phases. 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 de **machine learning** 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 à la diffusion de données.

Plusieurs projets sont réalisés au cours de l'année afin de mettre en pratique les divers enseignements dispensés au cours de cette année. Ils 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, ils sont aussi l'occasion de mettre en œuvre les compétences générales suivantes :

- Connaître les fondements des architectures distribuées, des réseaux aux systèmes répartis, des architectures logicielles big Data ;
- Avoir une ouverture sur les nouvelles technologies et savoir veiller à leurs évolutions ;
- Maîtriser les outils permettant l'extraction, l'intégration, l'analyse et la fouille de données (datamining);
- Maîtriser les outils permettant de manipuler de grands volumes de données et de créer des entrepôts de données ;
- Capacité à choisir et mettre en place le meilleur modèle de machine learning pour réaliser le traitement souhaité ;
- Capacité à concevoir et implémenter des systèmes de traitement de l'information ;
- Capacité à choisir, mettre en place et administrer la technologie de stockage la plus adaptée à un besoin précis (bases de données relationnelles, multidimensionnelles ou non relationnelles) ;
- Capacité à mobiliser les techniques de sécurisation des données à tous les niveaux, stockage, échange et diffusion ;
- Capacité à concevoir et mettre en œuvre un projet de développement d'applications informatiques (connaissances de base en architecture des systèmes, en réseaux, en sécurité, en méthodes de conduite de projet, méthodes de développement d'application).

	Volume horaire				
	Cours	**Ateliers**	**Projets**	**Total**	**Crédits**
UE0 Tronc commun					
Droit du Travail	3	6		9	0,5
Anglais	30			30	1,5
Sport		30		30	0
Total	**33**	**36**		**69**	**2**
UE1 Machine learning					
Machine learning	27	36		63	4
Régression pénalisée et sélection de modèles	9	6		15	1
Apprentissage statistique à grande échelle	6	9		15	1,5
Webmining et traitement du langage	9	18		27	2,5
Total	**51**	**69**		**120**	**9**
UE2 Développement d'application					
Génie logiciel	27	27		54	3
Indexation web	9	6		15	1
Projet web	12	15	9	36	2
Total	**48**	**48**	**9**	**105**	**6**
UE3 Big data					
Technologies sémantiques	6	9		15	1
Technologies NoSQL	12	3		15	1
Publication de données respectueuse de la vie privée	15	6		21	1
Total	**33**	**18**		**51**	**3**
UE4 Système et réseaux					
Réseaux et systèmes d'exploitation	15	6		21	1,5
Initiation à Unix	9	6		15	0
Systèmes Répartis	15	6		21	1
Sécurité des données	9	6		15	1
Grandes masses de données sur Cloud	12	12		24	1,5
Total	**60**	**36**		**96**	**5**
UE Projet de fin d'étude					
Projet méthodologique		9	27	36	2,5
Projet de fin d'étude		9	27	36	2,5
Data Challenge		12		12	0
Total	**0**	**30**	**54**	**84**	**5**
UE Projet professionnel et stages					
Stage de fin d'études					25
Stage d'application					5
Séminaires et projet professionnels	30			30	0
Total	**30**	**0**	**0**	**30**	**30**
TOTAL	**255**	**237**	**63**	**555**	**60**

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Descriptifs des enseignements communs

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE 0

UE : COURS D'OUVERTURE

Correspondant de l'UE : Ronan Le Saout
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 2
: Entre 50h et 60h

: 39h

Finalité de l'UE :

À la fin de cette UE, notamment grâce à l'enseignement de l'anglais, les élèves seront capables de mettre en œuvre les compétences linguistiques et culturelles qui facilitent la suivie des cours scientifiques dispensés en anglais ou d'autres langues, le travail dans un environnement professionnel international et la compréhension des normes culturelles dans les pays étrangers. À travers le cours de droit du travail, les élèves acquerront les connaissances dans une discipline autre que la statistique, l'économie et l'informatique nécessaires pour mieux appréhender le contexte juridique de l'entreprise. Cette UE vise également le développement des compétences transversales (*soft skills)* qui aideront les élèves à réussir les projets académiques de leur formation, à intégrer le marché du travail et à devenir les citoyens éclairés.

Structuration de l'UE :

L'UE 0 de la 3ème année se compose de deux matières obligatoires, l'anglais et le droit du travail, ainsi que le sport de manière optionnelle.

Compétences ou acquis d'apprentissage à l'issue de l'UE :

En anglais les élèves travaillent toutes les compétences linguistiques pour atteindre le niveau B2 du CECR et progresser vers un niveau C1. Lors des cours d'anglais et de l'aide au projet en anglais, les élèves développement également les connaissances liées au monde de l'entreprise et de la recherche ainsi que les compétences transversales *(soft skills).* Le cours du droit du travail permet aux étudiants d'identifier et comprendre certaines notions pratiques essentielles en gestion des ressources humaines en entreprise.

Les pré-requis de l'UE :

Aucun

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE 0 Cours d'ouverture

ANGLAIS

English

Enseignant : Divers intervenants (correspondant : Todd Donahue)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Répartition des enseignements
Langue d'enseignement : 1
: 40h

: 15h de cours, 15h d'aide au projet de fin d'études
: Anglais
Logiciels : Sans objet
Documents pédagogiques : Sous Moodle
Pré-requis : Aucun

Modalités d'évaluation :

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 (Cadre européen commun de référence pour les langues). Le CV et la lettre faite pour cet exercice seront évalués et feront partie de la note finale. L'anglais est également évalué à travers le rapport écrit et la soutenance orale du projet de fin d'études. Le niveau acquis apparaîtra sur le Supplément au diplôme. L'objectif de la CTI (Commission des Titres d'Ingénieur) pour tous les élèves ingénieurs est d'atteindre le niveau B2.

Acquis d'apprentissage (objectifs) :

- maîtriser une ou plusieurs langues étrangères
- savoir candidater et réussir un recrutement en langue anglaise
- contextualiser et prendre en compte les enjeux et les besoins de la société
- se connaître, s'auto-évaluer, gérer ses compétences, opérer ses choix professionnels
- s'intégrer et évoluer dans un groupe pour mener à bien un projet dans un contexte international et/ou pluriculturel
- savoir identifier les informations pertinentes, à les évaluer et à les exploiter

Principales notions abordées :

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

Références bibliographiques : Définies par chaque intervenant.

UE Cours d'ouverture

DROIT DU TRAVAIL

Work Law

Enseignant : Charlotte GRUNDMAN, Avocat au Barreau de Paris
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1
: 15h

: Cours : 3h • Atelier : 6h
: Français
Logiciels : Sans objet
Documents pédagogiques : Distribués pendant le cours
Pré-requis : Aucun

Modalités d'évaluation :

Exposé d'un cas pratique réalisé lors des TD.

Acquis d'apprentissage (objectifs) :

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.

Principales notions abordées :

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. Le cours commun (3 heures) traite des notions suivantes : Comprendre d'où l'on vient pour savoir où on va (introduction historique au droit du travail, les sources du droit du travail, ordre public absolu et ordre public social), les instances de contrôle du droit du travail, formation et exécution du contrat de travail, la rupture du contrat à durée indéterminée. Pour les TD, la première heure de cours sera consacrée à l'étude d'un chapitre (la modification du contrat de travail, le recrutement, les droits fondamentaux du salarié). 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.

UE Cours d'ouverture

SPORT

Sport

Enseignant : Divers intervenants (correspondant : Jullien Lepage)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel) : 0
: 30

Modalités d'évaluation :

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.

Acquis d'apprentissage (objectifs) :

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

Principales notions abordées :

Neuf 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.

UE 1

UE : APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING)

Correspondant de l'UE : Arthur Katossky
Nombre d'ECTS

Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 6 pour les filières ISTS, GDR et SSV, 7 pour GS et MQRM, 8 pour SID
: De 25h à 30h par ECTS

: 78h pour les filières ISTS, GDR et SSV, 102h pour GS et MQRM, 120h pour SID

Finalité de l'UE :

L'apprentissage automatique (machine-learning) est un paradigme essentiellement différent des approches statistiques exploratoires (statistiques au sens classique) ou explicatives (économétrie). Il vise un objectif de prédiction dans la continuité des méthodes d'apprentissage statistique supervisé introduites lors des premières années de la formation d'ingénieur. Largement utilisé dans l'ensemble des professions statistiques à l'heure actuelle (les métiers de la Data Science), l'apprentissage automatique est incontournable dans la formation de l'ingénieur statisticien et trouve de nombreuses applications: prédiction des cours basés à partir d'articles de presse en finance, détection de maladie par imagerie médicale en santé, recommandation de produits en marketing, compression d'images ou encore modèles de traitement du langage, toutes ces applications reposent sur les mêmes bases.

Structuration de l'UE :

L'UE se compose de 4 matières : apprentissage automatique (Machine-learning), régression pénalisées et régularisation, apprentissage statistique à grande échelle, traitement automatique de la langue et fouille du web (Natural language processing and webmining). L'ensemble de ces matières permettent de mettre en œuvre les techniques classiques, en développant un esprit critique sur leurs limites (sur-apprentissage, grande dimension, représentativité de l'échantillon) et en utilisant des données non structurées (texte, image…). Selon les filières de spécialisation, des séminaires complémentaires (systèmes de recommandation…) sont introduits.

Compétences ou acquis d'apprentissage à l'issue de l'UE :

Cette UE permet de maitriser des méthodes et des outils de l'ingénieur (identification, modélisation et résolution de problèmes même non familiers et incomplètement définis, l'utilisation des approches numériques et des outils informatiques, l'analyse et la conception de systèmes) en développant l'aptitude à étudier et résoudre des problèmes complexes, à concevoir et mettre en œuvre des projets de collecte et d'analyse d'informations et à concevoir et mettre en œuvre des algorithmes prédictifs de machine learning, s'intégrant dans une architecture informatique de données volumineuses (big data).

Les pré-requis de l'UE :

Modélisation statistique de 2ème année, méthodes d'optimisation et d'algorithmique, panorama du big data, aisance en R et Python.

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Machine Learning

APPRENTISSAGE AUTOMATIQUE

Machine Learning

Enseignant : Hong-Phuong DANG (Ensai), Romaric GAUDEL (Ensai), Fabien NAVARRO (Ensai) et Brigitte GELEIN (Ensai)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements

Langue d'enseignement : 2 (ISTS, GDR, SSV), 3 (GS, MQRM) ou 4 (SID)
: De 25h à 30h par ECTS

: Pour les 27h en filières ISTS, GDR et SSV, il y a 15h de cours et 12h d'ateliers. Pour les 24h complémentaires en GS et MQRM, il y a 6h de cours et 12h d'ateliers. Pour les 36h complémentaires en SID, il y a 12h de cours et 18h d'ateliers.
: Français
Logiciels : R et Python
Documents pédagogiques : supports de cours, bibliographie et fiches de TP
Pré-requis : R, Python, modélisation statistique, algèbre linéaire, optimisation de fonctions

Modalités d'évaluation :

Contrôle continu à la discrétion des intervenants, un QCM, compte-rendus de TP (1 en ISTS, GDR et SSV, 4 en GS et MQRM, 4 en SID)

Acquis d'apprentissage (objectifs) :

- 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

Principales notions abordées :

Un rappel des principes de l'apprentissage statistique et automatique sera effectué. L'ensemble des filières aborderont les réseaux de neurones (y compris deep learning), les méthodes d'agrégation (forêts aléatoires, bagging, boosting, stacking) et les séparateurs à vaste marge (Support Vector Machines). Les réseaux de neurones avancés seront abordés en GS, MQRM et SID, les systèmes de recommandation en MQRM et SID.

Références bibliographiques :

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

- 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

UE Machine Learning

REGRESSION PENALISEE ET SELECTION DE MODELES

Penalized problems and model selection

Enseignant : Cédric HERZET (INRIA) & Clément ELVIRA (INRIA)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1
: 30h

: Cours : 9h • Atelier : 6h
: Français
Logiciels : Python
Documents pédagogiques : Supports de cours, Supports de TP, Bibliographie
Pré-requis : Optimisation, Python, Algèbre

Modalités d'évaluation :

Un examen sur table de 2 heures avec questions de cours et résolution de problèmes, 1 compte-rendu de TP

Acquis d'apprentissage (objectifs) :

- identifier les cas d'apprentissage statistique et les problèmes inverses où la régularisation est utile : comprendre quels sont les motivations et les objectifs de la pénalisation
- connaître les modèles de régularisation les plus courants et savoir quelles caractéristiques de reconstruction ils favorisent choisir une régularisation parmi les méthodes les plus courantes et estimer un modèle régularisé par une méthode de descente de gradient
- connaître et comprendre les différents types de problèmes d'optimisation et les algorithmes qui permettent de les résoudre numériquement

Principales notions abordées :

- Ingrédients principaux des problèmes inverses et d'apprentissage statistique + exemples pratiques
- Motivations et objectifs de la régularisation
- Types de régularisation et fonctions de régularisation couramment rencontrées
- Caractérisation des problèmes pénalisés: existence de solution, unicité, conditions d'optimalité.
- Méthodes numériques de résolution de problèmes d'optimisation
- Conditions théoriques de reconstruction correcte

Références bibliographiques :

- Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.
- 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. Birkhaüser, 2013.
- D. P. Bertsekas. Nonlinear Programming. Athena Scientific, USA, 2003.

UE Machine Learning

APPRENTISSAGE A GRANDE ECHELLE

Large scale machine-learning

Enseignant : Romaric Gaudel (ENSAI)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1,5
: 25h

: Cours : 6h • Atelier : 9h
: Français
Logiciels : Python
Documents pédagogiques : support de cours, fiches de TP
Pré-requis : panorama du big data, machine-learning, optimisation

Modalités d'évaluation :

2 quizz en contrôle continu, 1 TP noté

Acquis d'apprentissage (objectifs) :

Le passage à des bases de données à grande échelle modifie certains usages en apprentissage statistiques. Nous en verrons quelques exemples, avec les justifications théoriques sous-jacentes.

- objectif : être en mesure de choisir des approches appropriées pour un problème donné. Nécessite de
- objectif : connaître le comportement en termes de taille de stockage et/ou de temps de calcul et/ou de qualité d'estimation des approches présentées.

Principales notions abordées :

- la descente de gradient stochastique
- données et modèles parcimonieux
- le calcul distribué

Références bibliographiques :

- Introduction to High Performance Computing for Scientists and Engineers, Georg Hager, Gerhard Wellein, CRC Press, 2010
- Introduction to High Performance Scientific Computing, Victor Eijkhout , Edmond Chow, Robert van de Geijn, 2014
- Bekkerman, Ron, Mikhail Bilenko, and John Langford. n.d. *Scaling up Machine Learning: Parallel and Distributed Approaches*.
- The MIT Press. n.d. “Large-Scale Kernel Machines.” Accessed September 1, 2020. https://mitpress.mit.edu/books/large-scale-kernel-machines.

UE Machine Learning

TRAITEMENT AUTOMATIQUE DU LANGAGE ET FOUILLE DU WEB

Webmining et NLP

Enseignant : Guillaume Gravier (Irisa)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1,5 sauf pour filière informatique (SID) : 2,5
: 38h sauf pour filière informatique (SID) : 56h

: Cours : 9h • Atelier : 12h (6h complémentaires en SID)
: Français
Logiciels : Python
Documents pédagogiques : Support de cours, Supports de TP
Pré-requis : Python, Machine Learning

Modalités d'évaluation :

Projet (de taille plus importante en SID)

Acquis d'apprentissage (objectifs) :

- collecter des données, extraire de l'information et apparier des sources textuelles
- choisir une méthode de traitement automatique de la langue pour une tâche classique (classification, analyse de sentiment, détection > d'entités…)
- se repérer parmi le foisonnement des modèles d'étude de la langue

Principales notions abordées :

1. What's natural language and its processing
2. The representation of words
3. The representation and classification of documents
4. Language modeling and contextual word embedding
5. Sentence-level tagging (token level tasks)
6. Sequence to sequence models and transformers
7. Overview of standard NLP tasks today

Références bibliographiques :

- Daniel Jurafsky, James H. Martin. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd edition, Prentice-Hall, 2009. Draft of the 3rd edition partly available at https://web.stanford.edu/~jurafsky/slp3.
- Yoav Goldberg. Neural Network Methods for Natural Language Processing. 2017. An earlier draft is freely available online at http://u.cs.biu.ac.il/~yogo/nnlp.pdf.
- Kevin Gimpel's lectures (Toyota Technological Institute at Chicago and UChicago) on Natural Language Processing (https://ttic.uchicago.edu/~kgimpel/teaching/31190-s18/index.html) and on Advanced Natural Language Processing (https://ttic.uchicago.edu/~kgimpel/teaching/31210-s19/index.html).

UE Projets

UE : PROJETS

Correspondant de l'UE : Arthur Katossky
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 5
: Entre 120h et 150h

: Suivis réguliers avec les encadrants

Finalité de l'UE :

Les projets sont l'occasion pour les étudiants de mettre en œuvre leurs connaissances acquises à l'ENSAI sur des cas d'études concrets. Ils visent à mettre en œuvre les outils et connaissance acquises en statistique, en informatique et en économie, dans une démarche de résolution de problèmes concrets type ingénieur.
Les projets de déclinent en deux versions: le projet académique, en langue anglaise, vise à approfondir une thématique centrée autour d'un ou plusieurs articles scientifiques ; le projet de fin d'études, plus appliqué, nécessairement sur des données issues d'une collecte, vise à proposer une solution pratique à une problématique générale proposée par une entreprise ou un laboratoire de recherche. À eux deux, ces projets couvrent toute l'étendue d'une démarche de développement: diagnostique d'un problème nouveau, lecture de la littérature scientifique sur le sujet, résolution d'un problème en respectant un compromis entre les règles de l'art d'une part, et les contraintes humaines, financières et techniques de l'autre. Ils permettent par ailleurs aux élèves de mesurer l'utilité de toutes les notions acquises au cours des trois années de formation.
Selon les filières, la réalisation d'un Data Challenge complète ces cas d'études concrets, à travers la réalisation d'un projet sur un temps court et des contraintes spécifiques.

Structuration de l'UE :

Projet méthodologique: approfondissement d'une démarche rigoureuse, à la pointe de la recherche scientifique, en langue anglaise ; constitue la partie théorique de recherche d'information dans une démarche de recherche et développement.
Projet de fin d'étude: approfondissement d'une démarche pratique, sachant composer avec des contraintes opposées, entre rigueur scientifique et nécessités pratiques ; constitue la partie implémentation dans une démarche de recherche et développement.
Data Challenge (optionnel, selon les filières) : rassembler sur une période très courte différentes équipes de profils variés afin de collaborer sur un projet.

Compétences ou acquis d'apprentissage à l'issue de l'UE :

Ces projets concluent la formation d'ingénieur de l'Ensai, et mobilisent un ensemble de compétences de l'ingénieur : capacité à trouver l'information pertinente, à faire une veille scientifique, à prendre en compte les enjeux de l'entreprise, à travailler dans un contexte international, tout en mobilisant des compétences techniques pour résoudre des problèmes complexes, et mener une démarche scientifique.

Les pré-requis de l'UE :

Méthodes de travail des projets de 1ère et 2ème année.

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Projets

PROJET MÉTHODOLOGIQUE

Methodological project

Enseignant : Divers intervenants
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 2,5
: Entre 60h et 75 h

: 9h d'ateliers, et suivis réguliers
: Anglais
Logiciels : /
Documents pédagogiques : /
Pré-requis : /

Modalités d'évaluation :

Le projet méthodologique consiste en la production d'un article de synthèse sur un sujet de recherche à choisir parmi un catalogue. L'évaluation tient compte de l'article rédigé et de la réalisation d'une soutenance.

Acquis d'apprentissage (objectifs) :

Les objectifs du projet méthodologique, et donc les compétences qui sont renforcées grâce à celui-ci, sont multiples:

- familiarisation avec la forme des productions académiques (articles notamment), en lecture comme en écriture
- capacité à faire une revue de littérature mélangeant ouvrages scientifiques et professionnels
- mise en œuvre d'une démarche scientifique rigoureuse
- prise de conscience des enjeux autour de la reproductibilité des résultats de recherche
- communication sur des sujets techniques

À cela s'ajoute les objectifs spécifiques à la production d'un travail technique en langue anglaise: mise en œuvre d'un projet complexe en langue anglaise, communication écrite et orale, acquisition d'un vocabulaire spécialisé, maîtrise de différents niveaux de langues en terme de style (oral vs. écrit) et de technicité (vulgarisation vs. spécialisation), mise en place de stratégies pour faire face à des difficultés linguistiques.

Principales notions abordées :

Travail de recherche en groupe suivi par un chercheur (env. 5 séances) et un professeur d'anglais (4 séances).

UE Projets

PROJET DE FIN D'ETUDES

Methodological project

Enseignant : Divers intervenants
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 2,5
: Entre 60h et 75 h

: 9h d'ateliers, et suivis réguliers
: Français
Logiciels : /
Documents pédagogiques : /
Pré-requis : /

Modalités d'évaluation :

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. Le projet est évalué à travers un rapport et une soutenance.

Acquis d'apprentissage (objectifs) :

Les objectifs du projet de fin d'études, et donc les compétences qui sont renforcées grâce à celui-ci, 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

.

Principales notions abordées :

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

UE Projets

DATA CHALLENGE

Data Challenge

Enseignant : Divers intervenants industriels (correspondante : Salima El Kolei)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : Pas d'attribution d'ECTS
: 2 journées

: 12h d'ateliers
: Français
Logiciels : /
Documents pédagogiques : /
Pré-requis : Méthodes de travail des projets, Compétences statistiques et informatiques de 3ème année

Modalités d'évaluation :

Les élèves participent au data challenge proposé à l'Ensai ouvert également aux élèves de deuxième année. Il n'y a pas d'évaluation.

Acquis d'apprentissage (objectifs) :

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 les compétences transversales acquises dans ce contexte opérationnel. Les compétences qui sont renforcées grâce à celui-ci sont multiples:

- 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…)

Principales notions abordées :

Travail en groupe sur un temps court.

UE 0

UE : PROJET PROFESSIONNEL ET STAGES

Correspondant de l'UE : Patrick Gandubert
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 30
: Travail en entreprise

: 30h (séminaires)

Finalité de l'UE :

Cette UE correspond à des temps pédagogiques en lien direct avec les entreprises. Les séminaires professionnels ont pour objectif de présenter aux étudiants diverses problématiques auxquelles ils seront confrontés dans leur environnement professionnel. Il permet d'apporter des compléments par rapport à certains cours, et fait le lien entre les enseignements et les applications pratiques qui en découlent. Le projet professionnel permet de préparer les étudiants à leur entrée dans la vie professionnelle et aux stages, il est réalisé sur la 2ème et 3ème année de formation. Des simulations d'entretien de recrutement sont organisées en 3e année. Elles sont assurées par des recruteurs d'entreprises et d'organisations partenaires de l'Ensai. Les stages (application en 2ème année, fin d'études en 3ème année) permettent aux élèves de mettre en pratique les enseignements de mathématiques appliquées, d'informatique et d'économie dans un cadre professionnel. Le stage de fin d'études, d'une durée de 20 semaines minimum, vise à appliquer les enseignements de 3ème année et à acquérir de l'expérience pour assurer la transition vers l'emploi. Il constitue une étape essentielle de mise en situation professionnelle pour le futur ingénieur qui dispose à ce stade de l'ensemble des bagages techniques de la formation.

Structuration de l'UE :

Le stage de fin d'études constitue la majeure partie de l'évaluation de cette UE (25 ECTS). L'Ensai exige une forte adéquation entre le contenu du stage et la filière de spécialisation de 3e année. Il fait l'objet d'une procédure de validation par le responsable de filière et par le département des relations avec les entreprises. L'évaluation tient compte de la capacité d'intégration de l'étudiant dans l'entreprise, ses capacités d'initiative et de satisfaction au regard des objectifs du stage, et de la qualité du rapport et de la soutenance réalisée devant un jury composé d'un président, d'un vice-président, tous les deux issus du monde de l'entreprise et d'un permanent de l'école. Le stage d'application de 2ème année est pris en compte dans cette UE (5 ECTS). Les séminaires professionnels ne sont pas évalués.

Compétences ou acquis d'apprentissage à l'issue de l'UE :

Le stage de fin d'études (et l'UE) comprend un objectif technique - il s'agit de répondre à la commande, à la problématique inscrite dans le thème du stage à l'aide des connaissances acquises - et un objectif professionnel – il s'agit de parfaire la connaissance du monde du travail, de développer des capacités relationnelles et d'adopter une démarche d'insertion dans le monde professionnel.

Les pré-requis de l'UE :

Aucun

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Descriptifs des enseignements de la filière

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Spécifiques filière SID

UE SPÉCIFIQUES FILIÈRE SID

Correspondant de l'UE : Hong-Phuong Dang et Rémi Pépin
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 14
: De 25 à 30h par ECTS

: 285h

Finalité des UE :

Ces UE visent à renforcer les compétences informatiques des élèves ingénieurs, pour s'orienter vers des métiers de la Data Science directement en lien avec le génie logiciel, l'ingénierie des données ou l'inclusion de méthodes d'apprentissage statistique dans des architectures big data.

Structuration de l'UE :

La filière SID inclut 3 UE spécifiques : développement d'application, big data et système et réseaux.

Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Connaître les fondements des architectures distribuées, des réseaux aux systèmes répartis, des architectures logicielles big Data ;
- Avoir une ouverture sur les nouvelles technologies et savoir veiller à leurs évolutions ;
- Maîtriser les outils permettant l'extraction, l'intégration, l'analyse et la fouille de données (datamining);
- Maîtriser les outils permettant de manipuler de grands volumes de données et de créer des entrepôts de données ;
- Capacité à choisir, mettre en place et administrer la technologie de stockage la plus adaptée à un besoin précis (bases de données relationnelles, multidimensionnelles ou non relationnelles) ;
- Capacité à mobiliser les techniques de sécurisation des données à tous les niveaux, stockage, échange et diffusion ;
- Capacité à concevoir et mettre en œuvre un projet de développement d'applications informatiques (connaissances de base en architecture des systèmes, en réseaux, en sécurité, en méthodes de conduite de projet, méthodes de développement d'application).

UE2 – Développement d'application

Génie Logiciel

Software engineering

Cours : 27h • Atelier : 27h

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

Correspondant : Rémi PEPIN

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**
   1. Introduction au génie logiciel et bonne pratiques de conception.
   2. Architecture logicielle et modèle en couche, Exemple sur GWT.
   3. Principaux patrons de conception, principe et mise en œuvre en Java.
   4. Le test logiciel et l'ingénierie des langages.
   5. 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 : Rémi PEPIN

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.

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

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	:	Rémi PEPIN

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

Français

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE3 – Big Data

Technologies Sémantiques

Semantic Technologies

Cours : 6h • Atelier : 9h

Enseignants : Sébastien FERRE (IRISA)

Correspondant : Rémi PEPIN

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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE3 – Big Data

Technologies NoSQL

NoSQL Technolgies

Cours : 12h • Atelier : 3h

Enseignant : David GROSS-AMBLARD (IRISA)

Correspondant : Rémi PEPIN

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

Objectif pédagogique

Ce cours vise 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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE – Big Data

Publication de données respectueuse de la vie privée

Privacy-preserving data publishing

Séminaire : 9h

Enseignant : Tristan ALLARD (Univ. Rennes 1)

Correspondant : Rémi PEPIN

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 :
QCM en fin de présentation

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:

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

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 : Rémi PEPIN

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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE4 – Systèmes et Réseaux

Initiation à Unix

Networks and Systems

Cours : 9h • Atelier : 6h

Enseignant	:	François Xavier BRU (Orange Consulting)
Correspondant	:	Rémi PEPIN

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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE4 – Systèmes et Réseaux

Systèmes Répartis

Networks

Cours : 15h • Atelier : 6h

Enseignant : David FREY & George GIAKKOUPIS (Inria Rennes)

Correspondant : Rémi PEPIN

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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE4 – Systèmes et Réseaux

Sécurité des données

Data Security

Cours : 9h • Atelier : 6h

Enseignant : Franck LANDELLE (DGA MI)

Correspondant : Rémi PEPIN

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 :

Aucun

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 : Rémi PEPIN

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