Compétences fonctionnelles

- **Data Science**, **Data Engineering** (Mise en production des algorithmes d'intelligence artificielle)
- **Conception de Dashboard**
- Machine Learning & Deep Learning
- Computer Vision, NLP
- Etat de l'art sur l'intelligence artificielle
- Formation sur la Data Science

Compétences techniques

- Langages : **PYTHON (expert)**, **R**, Matlab, C++, **JAVA**, Node.js
- Frameworks : Tensorflow, Cmusphinx, Caffe, Theano, Keras, SPARK, KAFKA, Cuda, Nutch, Scikit-learn
- Bases de données : **SQL**, Mongodb, Cassandra, **HADOOP**, Firebase, Cosmos DB, Neo4j, POSTGRESQL
- Systèmes : Windows, Unix, Linux
- **Data vizualisation** : **D3.js**
- Méthodologie projet : Agile **SCRUM**
- Outils : Confluence, Slack, JIRA**, GIT HUB**
- Développement Web : **REST**
- Cloud Computing : **MS AZURE, MS AZURE DATABRICKS, MS Cloud AZURE , AWS**
- Virtualisation : **DOCKER**

Formation

- **ENSAE** : Ingénieur en statistique (2016)
- **Ecole Normal Supérieur** (ENS) : Master MVA (Mathématiques, Vision, Apprentissage)
- Prépa MP / MPSI – Lycée Henry IV

Langues

Anglais : bilingue

Data Scientist – Data Engineer

Eléments de parcours

Biogen – Data Analyst, Paris (Fév. 2020 – Avr.2020)

Projet : Konectom
*Konectom est une **application mobile** de Biogen. Elle consiste en un medical device permettant de suivre l'évolution de maladies cérébrales : SMA, MS, … L'application évite aux patients de faire certains tests chez le médecin dans le cadre de leur suivi.*
Client : Equipe scientifique

Contexte / Objectif : Le projet s'inscrit dans l'intégration d'un produit d'une start-up (Ad Scientiam) dans les produits de Biogen permettant de combiner l'innovation avec la force de frappe de Biogen au niveau des Clinical trials. L'équipe Data est chargée d'analyser les données d'utilisation de l'application. L'objectif est la mise de place de features optimaux permettant le suivi de la progression de la maladie chez le patient. Premier Data Scientist / Data Analyst recruté au sein de l'équipe Data.

Environnement travail / fonctionnel : Equipe de 10 personnes en Agile Scrum.

Contraintes : Le Coronavirus qui a mis un stop aux essais cliniques du produit. Aucune infrastructure d'analyse de données au début de la mission.

Principales réalisations :

- Suivi d'une entreprise tierce pour la mise en place du back-end de la plateforme de Data Science :
  - Analyse des besoins utilisateurs (Data Scientists)
  - Revue de l'architecture back-end proposée sur le cloud AWS (Amazon Web Services) avec l'entreprise tierce - Cycle de vie des données : base de données firebase, document Db, lambda functions, notebooks, **intégration Git**, outils d'analytics
  - **Définition des formats de données** utilisées sur la plateforme de Data Science et récupérées sur les devices
  - Définition des variables utilisées à partir des réunions scientifiques

- Mise en place d'une première librairie data science avec les fonctionnalités CI/CD sur le GitLab de l'équipe Data Engineering :
  - Création d'un algorithme état de l'art de step detection sur des **séries temporelles de données d'accélération et de rotation** à partir d'un papier de recherche pour le U-turn test et mise en place d'une démo (visualisation des résultats de l'algorithme en temps réel) permettant de démontrer l'algorithme à l'équipe
  - Lancement d'une étude de reliability avec **R** afin de visualiser la qualité des features et de leur **stabilité temporelle.**

Environnement technique : AWS, **Python**, Docker, **R**, mongo Db, Signal processing

AXA Group Operations – Data engineer, Lausanne (Nov. 2019 – Janv. 2020)

Projet : Healthcare bot

Au sein d'AXA REV (Research Engineering Vision) qui est le pôle innovation de l'IT au niveau groupe d'AXA. Les équipes d'AXA REV travaillent sur des projets à haut potentiel et placés très haut sur la value chain. Elle collabore avec les meilleurs chercheurs du monde (Stanford, Ecole Polytechnique Fédérale de Lausanne, …)

Client : N+2 de l'équipe AXA REV

Contexte / Objectif : Le projet consiste en la mise en place d'un Chatbot médical. L'utilisateur peut recevoir des diagnostics et suggestions de médecins en échangeant avec le Chatbot.

Environnement travail / fonctionnel : Equipe de 10 personnes réparties à Lausanne, Paris et Barcelone et un Chef de Projet Agile.

Contraintes :

- Contrainte de temps : Déploiement du bot en production mi-janvier.
- Contraintes légales, de sécurité, de robustesse, d'image de marque, ...

Principales réalisations :

- **Aide à la définition et mise en place de l'architecture** back-end (fonctions Azure, serveur FHIR, transformation de données grâce à Typescript, Spark…) du chatbot enregistrant les interactions utilisateurs dans une base de données graph sur **Azure** Cosmos DB regroupant les données anonymisées d'utilisation du bot.
- Utilisation de Docker pour tester les fonctions en local.
- Travaux effectués sur la robustesse du code avant la mise en production et l'intégration de test pour l'intégration continue.
- **Code review** avec le project manager en s'assurant des **best practices Python**.

Environnement technique : Cloud Azure, fonctions **Azure**, Trigger, procédures **Azure**, Databricks sur **Azure**, l'API REST de Spark, Neo4j, Standard FHIR, Typescript, **Javascript**, VS code, **GitHub**, **Python** Cosmos DB

Servier – Computer Vision Data Scientist (Juin 2019 - Sept. 2019)

Projet : Détection de cancer sur les images histopathologiques

L'équipe Data Science fait partie du PEX MVD (Pôle d'Expertise 'Modélisation et Valorisation des Données') intégré à IRIS (Institut de Recherches Internationales Servier). Elle travaille sur des problématiques de Computer Vision, NLP, Séries temporelles, anonymisations appliquées au domaine médical.

Client : Project Manager Data Science

Contexte / Objectif : Détection de cancer avec des algorithmes de Deep Learning pour identifier les différents types de cancer. **Projet de classification de séries temporelles.**

Environnement travail / fonctionnel : Equipe de 6 personnes dont 2 Chefs de Projet.

Contraintes : Contrainte de temps pour une mise en place des algorithmes avant fin septembre. Limitation au niveau des librairies Deep Learning : utilisation des librairies et versions qui sont installées sur le HPC (>10 GPU, plusieurs Po de stockage). Nombre assez faible d'images par cancer.

Principales réalisations :

- Etat de l'art d'algorithmes de traitement d'images 3D dans le but de les appliquer sur des scans IRM du genou pour détecter la progression de l'arthrose (3D CNN, …)
- Mise en place d'un algorithme de Computer Vision (ResNet) avec du Transfer Learning pour la détermination du type de cancer (Prostate, Colon, sein, poumon). Détermination des étapes de nettoyage et mise en forme des images
- **Extraction de features de séries temporelles** de recrutement clinique pour faire de la **classification (K-means)** au niveau du site-étude afin d'avoir un prior pour entrainer un modèle bayésien

Environnement technique : HPC, Tensorflow, Selenium, **Python**

Samsung – Machine Learning Scientist & Engineer (Oct. 2018 – Juin 2019)
Projet : ASR (Automatic Speech Recognition)
L'équipe ASR s'occupe de la transcription de la voix en texte. Les performances de l'équipe ASR a une influence directe sur la reconnaissance vocale de Bixby (Assistant virtuel) qui est installé sur le téléphone des particuliers. L'équipe est divisée en plusieurs entités : Language Modeling, ITN, Wake up, G2P.

Client : Manager France Bixby

Contexte / Objectif : L'objectif est d'améliorer la reconnaissance vocale de Bixby sur les devices Samsung des utilisateurs.

Environnement travail / fonctionnel : Equipe de 10 personnes en Agile **Scrum** dont un Team Leader.

Contraintes :
- Encadrement du projet par Samsung Pologne qui a mis en place l'architecture permettant de sous-traiter les langues Européennes
- Contraintes de temps avec la sortie du produit en novembre 2018
- Contrainte de performances pour l'acquisition de nouveaux clients
- Contraintes légales pour le crawling de sites web.

Principales réalisations :
- Crawling de sites web pour entrainer le modèle de langage. Utilisation de Rotating Proxies, BeautifulSoup, Headless Chrome, Selenium (utilisé par exemple pour des pages à défilement infini sur le forum Quora), VPN, Apache Nutch (par exemple pour crawler le monde), Wikipédia, Français facile, Opus.... Tests de différents agencements de textes de language naturels pour améliorer la perplexité
- Mise en place de slot dictionaries
- Automatisation de tâches de correction de corpus
- Phonétisation automatique via LSTM

Environnement technique : Regular expression, Nutch, **Python**, CMU Sphinx

SoftatHome – Data Scientist (Mars. 2018 – Oct. 2018)
Projet : Eyes'on
Surveiller et optimiser l'utilisation du Wifi des utilisateurs.

Client : Directeur de l'équipe projet Eyes'on.

Contexte / Objectif : Softathome est spécialisé dans les logiciels des box wifi. Le projet Eyes'on se focalise sur la récupération des données d'utilisation des box wifi (RSSI, force du signal, ...) pour les stocker dans une base de données. Ces données sont ensuite analysées pour apporter de la valeur.

Environnement travail / fonctionnel : Equipe de 10 personnes de Data Engineers, Data Scientists et Data Analysts – Agile **Scrum** - utilisation de Jira / Confluence.

Contraintes : Utilisation du Cloud Orange et de 2 Alienware

Principales réalisations :
- **Analyse de données massives** (via **Kafka**) des clients de Softathome (O2, Orange) afin d'aider les opérateurs à prendre des décisions plus rapidement
- Mise en place d'un environnement **Spark** pour réaliser du calcul parallèle. Un cluster S3 local a aussi été mis en place. Transfert des données de la base de données Cassandra vers Amazon S3
- **Analyse statistique des données de séries temporelles** wifi et **data vizualisation**
- **Proposition d'une architecture** intégrant Spark avec Kafka pour faire de la prédiction quasi en temps-réel.
- **Evangélisation de l'équipe au niveau de l'utilité et des use-cases** de Spark pour le traitement de données massives.

Environnement technique : Cassandra, **Kafka**, Spark, **Python**, Unix, **Git**, Jira, Amazon S3, Orange Cloud

Dassault Systèmes – Data Scientist (Nov. 2016 – Mars 2018)
Projet : Systèmes de recommandation
Etat de l'art des systèmes de recommandation

Client : Manager Data Science de l'équipe Recherche

Contexte / Objectif : La *3D Experience* plateforme de Dassault Systèmes est une marketplace de produits 3D. L'objectif du projet est de **réaliser un PoC** et un état de l'art dans le but futur d'intégrer un système de recommandation liant les utilisateurs de produits 3D et les vendeurs de produits 3D.

Environnement travail / fonctionnel : Equipe de Recherche composée de 10 personnes spécialisées en Machine Learning. Méthodologie projet Cycle en V avec état d'avancement du projet. **Points de suivi hebdomadaire avec l'équipe Data Science. Mise à jour des objectifs annuels et du réalisé avec le management.**

Contraintes : Peu de contraintes étant un projet de recherche.

Principales réalisations :

- Implémentation d'un algorithme basé sur des graphes combiné avec du Tf-IDF, les données étant en faible quantité. Cet algorithme avait donné les meilleurs résultats au challenge 2016 de recommandation de Dassault Systèmes.
- Etat de l'art des systèmes de recommandation
- Modification de l'architecture du Denoising Autoencoder pour qu'il puisse prendre en compte les features utilisateurs et les features items.
- Implémentation d'un système d'interprétation permettant d'expliquer les recommandations du Denoising Autoencoder (LIME algorithm)
- **Conception d'un Dashboard Flask** (à destination des commerciaux équipe marketing): mise en place en place d'un serveur interne pour valoriser les résultats des systèmes de recommandations. Intégration des **visualisations D3.js au dashboard** et **présentation au ComEx** et aux commerciaux.
- Transfert de la technologie de Denoising Autoencoder vers **Spark** pour intégration

Environnement technique :
Python, Réseaux neuronaux, **Machine Learning**, Caffe, Theano, Cuda, UNIX, Flask, **D3.js**, Modèles bayésiens, Spark, multiprocessing, Camtasia, machines virtuelles, **MongoDB. Git**, Docker.

Dassault Systèmes – Data Scientist (Mai. 2016 – Oct. 2018)
Projet : Text classification
Classification de queries des utilisateurs sur la knowledge base

Contexte / Objectif : Le data scientist devait analyser le parcours client sur la knowledge base afin de faire ressortir des patterns. L'objectif étant de spécifier des clusters et thresholds qui seraient intégrés aux outils du support client afin de lancer une action mieux ciblée.

Environnement travail / fonctionnel : Equipe Knowledge Management de 5 personnes qui a la responsabilité de l'amélioration continue de l'expérience utilisateur de la Knowledge base de Dassault Systèmes.

Contraintes : Rédiger un rapport de stage et **réaliser un PoC** qui permettrait **d'améliorer l'expérience utilisateur.**

Principales réalisations :

- Analyse des logs de recherche des clients de Dassault sur la knowledge base dans le but de faire du clustering (objectif : Détecter les différents sujets qui intéressent les utilisateurs) et détecter les anomalies pour mettre en place un système de feedback automatisé au support client.
- **Mise en place d'algorithmes de clustering** (K-means, algorithmes basés sur la théorie de graphes, ...) sur des features extraits grâce aux méthodes de NLP (tf-idf, word embeddings, ngrams, ...) afin de classifier les requêtes utilisateurs.
- Aide à la mise en place d'une démo pour la présenter au ComEx.
- Collaborer avec un membre de l'équipe recherche pour la classification d'images de la Knowledge base pour améliorer le moteur de recherche (CNN classique).

Environnement technique : Python, NLTK, R, Theano