# Offre de stage 2018 : informatique et statistique

## Sujet

Panorama et évaluation des offres MLaaS (Machine Learning as a Service)

## Contexte

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long termes. Dans ce cadre, le département Innovation Commerciale Analyse des Marchés et de leur Environnement (ICAME) est un département pluridisciplinaire (sciences de l'ingénieur, sciences humaines et sociales) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils de relation client aux directions opérationnelles du groupe EDF. Au sein de ce département , le poste sera rattaché au groupe «Statistiques et Outils d'Aide à la Décision » (SOAD) qui compte une vingtaine d'ingénieurs chercheurs spécialisés en data mining, informatique décisionnelle et data science ayant pour mission de construire et mettre en œuvre les méthodes d'analyse, de fouille et d'enrichissement de données volumineuses d'origines multiples, structurées ou complexes. Le stage sera co-encadré par un ingénieur-chercheur du département PRISME (Performance, Risque Industriel, Surveillance et Maintenance pour l'Exploitation). Le département PRISME a pour vocation d'apporter des solutions innovantes pour une exploitation toujours plus performance des moyens de production du Groupe EDF.

La **datas science** (ou science des données) est un domaine interdisciplinaire qui vise l'extraction de connaissance des données à partir de l'application des théories et des techniques tirées de plusieurs domaines (mathématique, statistique, technologie de l'information, etc.). Le **machine learning** (ou apprentissage automatique) est l'une des techniques de la data science. Le mot « apprentissage » signifie que les méthodes dépendent des données (utilisées comme un ensemble d'apprentissage) pour paramétrer un modèle ou algorithme spécifique (ex : prévision, régression, classification, etc.). Afin de préparer le Groupe EDF aux enjeux stratégiques du Big Data et de a Data Science, une équipe de la R&D travaille activement sur ces sujets. Nous recherchons un(e) stagiaire pour contribuer aux travaux du projet R&D Data Impulse.

## Objectifs

Plusieurs offres packagées de type **MLaaS (Machine Learning as a Service)** sont proposées aujourd'hui par des sociétés externes (GE, Microsoft, Amazon, Google, etc.). Dans le cadre du suivi des évolutions des technologies liées au machine learning, nous souhaitons élargir nos études ponctuelles à une évaluation méthodologique et étendue des offres MLaaS consolidées mais aussi en émergence. L'objectif de ce stage est d'évaluer et comparer plusieurs offres MLaaS afin d'en identifier le potentiel et les limites en fonction des besoins métiers d'EDF. De plus, il s'agira de mener des tests permettant d'estimer la performance de l'outil, sa prise en main, l'import/export de données, les méthodes proposées, son intégration dans le SI, etc. Des briques plus spécifiques pourront également être testées (i.e. connecteurs spécifiques, inclusion d'algorithmes développés en interne, etc.).

Ce stage se décomposera en 3 parties :
- **Panorama** des offres MLaaS: Il s'agira de mener un état de l'art des offres existants (en open source ou propriétaires) et de repérer celles présentant un intérêt pour les activités d'EDF.
- **Evaluation** : A partir des solutions repérées, il s'agira d'évaluer une sélection d'outils jugés pertinents par rapport aux besoins d'EDF. Cette évaluation sera basée sur un ensemble de critères et des use-cases définis en accord avec les encadrants et inclut des tests sur une volumétrie représentative de données (simulées ou réelles) et portera sur différentes fonctionnalités de l'outil (import/export de données, méthodes d'analyse, publication et visualisation de données, etc.)
- **Comparaison** des résultats avec positionnement des offres entre elles (niveau de maturité, coûts et mode de facturation, support proposé, etc.)

**Profil recherché :**

- Etudiant(e) en master 2 ou troisième année d'école d'ingénieur
- Programmation (python, java, R, javascript, php, etc.) en environnement Linux et Windows
- Connaissance des méthodes mathématiques, statistiques et de machine learning
- Connaissance des environnements Big Data et cloud serait un plus
- Connaissance sur le secteur de l'énergie serait un plus

- Curieux(/se), ingénieux(/se) et motivé(e) pour le domaine de la recherche appliquée

# Informations pratiques

**Unité d'accueil :** Groupe SOAD (Statistique et Outils d'Aide à la Décision), département ICAME d'EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.
Le stage sera co-encadré par deux ingénieurs-chercheurs des départements ICAME et PRISME.

**Début du stage souhaité :** le plus tôt possible en 2018.

**Transmettre par mail un CV, une lettre de motivation et les bulletins de notes à** : jean-paul.le@edf.fr (Département ICAME) et christophe.denis@edf.fr (Département PRISME).

Ci-dessous des exemples de travaux en Data Science/Machine Learning publiés par des contributeurs du projet R&D Data Impulse :


- **A Data Lake and a Data Lab to Optimize Operations and Safety Within a Nuclear Fleet.** Marie-Luce Picard, Jean-Marc Rangod, Christophe Salperwyck. *Hadoop Summit 2016*, California, USA, June 2016: http://fr.slideshare.net/HadoopSummit/a-data-lake-and-a-data-lab-to-optimize-operations-and-safety-within-a-nuclear-fleet
- **Exploring Titan and Spark GraphX for Analyzing Time-Varying Electrical Networks.** Guillaume GERMAINE, Thomas Vial, *Hadoop Summit 2016*, Dublin. http://fr.slideshare.net/HadoopSummit/exploring-titan-and-spark-graphx-for-analyzing-timevarying-electrical-networks. Vidéo: https://www.youtube.com/watch?v=Xk8UPECiMSw
- **CourboSpark: Decision Tree for Time-series on Spark.** Christophe Salperwyck, Simon Maby, Jérôme Cubillé, Matthieu Lagacherie, *Hadoop Summit 2015*, Dublin, https://speakerdeck.com/simonmaby/courbospark-decision-tree-for-time-series-on-spark. Vidéo: https://www.youtube.com/watch?v=GNtU-kVL5xI
- **Computing Data Quality Indicators on Big Data Stream Using a CEP.** Wenlu Yang, Alzennyr Gomes Da Silva, Marie-Luce Picard, *IEEE Xplore - IWCIM 2015,* Prague, Novembre 2015. https://tel.archives-ouvertes.fr/LIP6/hal-01367862v1
- **Real-time energy data-analytics with Storm.** Rémy Saissy, Marie-Luce Picard, Charles Bernard, Bruno Jacquin, Simon Maby, Benoît Grossin, *Hadoop Summit 2014*, Californie, USA, 2014. http://fr.slideshare.net/Hadoop_Summit/t-525p212picard
- **HETA: Hadoop environment for text analysis.** Vincent Nicolas, Alzennyr Gomes da Silva, Marie-Luce Picard, *IWCIM (International Workshop on Computational Intelligence for Multimedia Understanding*), IEEExplorer, 2014, 10.1109/IWCIM.2014.7008803
- **Smart Metering x Hadoop x Frost: A Smart Elephant Enabling Massive Time Series Analysis.** Benoît Grossin, Marie-Luce Picard, *Hadoop Summit Europe 2013*, Amsterdam, Mars 2013. http://hadoopsummit.org/amsterdam/
- **Searching time-series with Hadoop in an electric power company.** Alice Bérard, Georges Hébrail, *BigMine Workshop*, KDD2013, Chicago, August 2013. http://bigdata-mining.org/
- **Simulation and forecasting electricity demand at scale.** Alexis Bondu, Yannig Goude, Marie-Luce Picard, Pascal Pompey, Mathieu Sinn, *European Utility Week*, Amsterdam, October 2013.http://www.european-utility-week.com/
- **Empower agile BI & analytics for utilities with a total data approach.** Marie-Luce Picard, Bruno Jacquin, *Teradata Partners Conference*, Dallas, October 2013. http://www.teradata-partners.com
- **A proof of concept with Hadoop: storage and analytics of electrical time-series.** Marie-Luce Picard, Bruno Jacquin, *Hadoop Summit 2012*, Californie, USA, 2012. http://www.slideshare.net/Hadoop_Summit/proof-of-concent-with-hadoop
- **Massive Smart Meter Data Storage and Processing on top of Hadoop**. Leeley D. P. dos Santos, Alzennyr G. da Silva, Bruno Jacquin, Marie-Luce Picard, David Worms,Charles Bernard. *Workshop Big Data 2012*, Conférence VLDB (Very Large Data Bases), Istanbul, Turquie, 2012. http://www.cse.buffalo.edu/faculty/tkosar/bigdata2012/program.php