Proposition de recherche doctorale

# Theory and practice of scalable machine learning algorithms

**Mots clés :**

- **Directeur de thèse** : pietro MICHIARDI
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire de recherche d'EURECOM
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

## Résumé du projet de recherche (Langue 1)

The amount of data created each second in our world is exploding. E-commerce, Internet security and financial applications, billing and customer services – to name only a few examples – will continue to fuel exponential growth of large pools of data that can be captured, communicated, aggregated, stored, and analyzed. As companies and organizations go about their business and interact with individuals, they are generating a tremendous amount of digital footprints, i.e., raw, unstructured data – for example log files – that are created as a by-product of other activities. The use of these huge quantities of data is considered today as a key basis of competition and growth: companies failing to develop their analysis capabilities will fail to understand and leverage the big picture hidden in the data, and hence fall behind. The current state-of-the-art already offers a set of approaches to tackle such large-scale data processing problems, like commercial databases (Oracle Big Data), public cloud services (Amazon Elastic MapReduce) and open-source projects (Hadoop). Nevertheless, designing scalable machine learning algorithms, that are able to discover compelling knowledge from these huge amounts of data, remains a hard problem. The high complexity of above mentioned execution frameworks, makes the design of such efficient algorithms complicated. Moreover, the optimization of these algorithms requires to understand the cost of these algorithms, which is also quite challenging. Finally, these systems make the implementation of even simple algorithms intricated. For example, implementations of even simple clustering algorithms that are largely used in many fields are inefficient and do not make an appropriate use of the underlying system resources (see for example the Mahout project). Therefore, the goal of this Thesis will be to develop highly scalable, optimized and reusable machine learning algorithms to process and interact with large amounts of data. The Thesis will not only focus on algorithm design, but also on the understanding and modelisation of the cost and bottlenecks of these algorithms. More generally, the Thesis should study the global purpose of these algorithms: does processing more data leverage more valuable knowledge, and do complicated distributed algorithms provide benefits compared to more simple algorithms?

## Résumé du projet de recherche (Langue 2)

Théorie et pratique des algorithmes de machine learning évolutifs La quantité de données produites chaque seconde dans le monde est en pleine explosion. Les applications d'e-commerce, de sécurité ou financières, les services à la clientèle ou de facturation, pour n'en citer que quelques-uns, vont continuer à alimenter la croissance exponentielle des énormes réservoirs de données qui peuvent être capturées, assemblées, stockées, et analysées. Lorsque les entreprises et autres organisations vaquent à leurs occupations et interagissent avec leurs clients ou utilisateurs, elles génèrent une énorme empreinte numérique, c'est à dire des données brutes, non structurées (par exemple les fichiers journaux) qui sont créées en tant que sous-produit d'autres activités. L'utilisation de ces énormes quantités de données est actuellement considérée comme une clé de la compétitivité et de la croissance: les entreprises incapables de développer les capacités d'analyse seront incapables de comprendre et d'utiliser les tendances cachées dans leurs données, et se retrouveront immanquablement à la traîne. L'état de l'art actuel offre déjà une série d'approches pour s'attaquer à de tels problèmes d'analyse de données à grande échelle, comme des bases de données commerciales (Oracle Big Data), des services de cloud publics (Amazon Elastic MapReduce) et des projets open-source (Hadoop). Néanmoins, concevoir des algorithmes de machine learning évolutifs, capables d'extraire des connaissances intéressantes et exploitables de ces énormes quantités de données, reste un problème difficile. La complexité élevée des fameworks d'exécution mentionnés ci-dessus rend compliquée l'élaboration d'algorithmes efficaces. De plus, l'optimisation de ces algorithmes nécessite une compréhension précise, et difficile, du coût de ces algorithmes. Enfin, ces systèmes rendent l'implémentation d'algorithmes, même simple, très complexe. Par exemple, les implémentation actuelles de simples algorithmes de clustering, qui sont régulièrement utilisés dans de nombreux domaines, sont inefficientes et n'utilisent pas efficacement les ressources disponibles dans le système sous-jacent (c'est par exemple le cas dans le projet Mahout). C'est pourquoi le but de cette thèse sera de développer des algorithmes de machine learning qui soient à la fois évolutifs, optimisés, et réutilisables, afin de traiter et d'interagir avec de grandes quantités de données. La thèse ne se concentrera pas sur la conception des algorithmes, mais également sur la compréhension et la modélisation de leurs coûts et goulots d'étranglement. De façon plus générale, cette thèse devra étudier le but global de ces algorithmes: analyser de plus grandes quantités de données permet-il effectivement d'extraire une connaissance plus intéressante des phénomènes cachés dans les données, et dans quelle mesure ces algorithmes complexes apportent ils un avantage concurrentiel par rapport aux algorithmes plus simples?