PhD Proposal 2017

School: Ecole Centrale de Lyon	
Laboratory: LIRIS	**Web site: http://liris.cnrs.fr/**
Team: SICAL / IMAGINE	**Head of the team: Jean-Charles Marty (SICAL), Veronique Eglin (IMAGINE)**
Supervisor: Romain Vuillemot **Stéphane Derrode**	**Email: romain.vuillemot@ec-lyon.fr** **stephane.derrode@ec-lyon.fr**
Collaboration with other partner during this PhD: **In France:**	**In China:**

Title: Human in the Loop to Harness Big Data: Application to Predictive Data Analytics
Scientific field: Computer Science
Key words: Big data, predictive models, visual analytics, human in the loop, information visualization

Details for the subject:

Background, Context:

Big data models are now pervasive and already have revolutionized many fields, impacting our day-to-day life. In healthcare, automated diagnostics now challenge traditional doctors in accuracy. In transportation, driverless rides will soon become a commodity with better safety than human drivers. In marketing, consumers segments can be refined, be more accurate and anticipate consumers' needs in real-time when fed with real-time, geo-located data. Such rapid advances have been driven by larger datasets and computational infrastructures, but also with classes of models (such as convolutional networks) that have become very efficient to quickly tackle complex problems.

However, little attention has been put on the human aspect of those big data models: there is indeed not much understanding for humans of what the models actually do, and what they are made of as they often are the combination of already complex, obscure models. Models are also increasingly difficult to configure and require many learning phases of trial and errors to find the best set of parameters or datasets to train them [Sedlmair14, Mühlbacher14]. Some companies even admit they can't really control or explain outcomes; they just try to "tame" their models and spend days understanding unexpected behaviors. Thus humans remain involved to configure those models as we (still) have better qualitative judgment and can figure out novel classes of problem never though of by the model, or which didn't exist in the training data [Pretorius11, Cancino12]. Finally, there is a growing need in models acceptability, both by experts to understand them, but also in customers or any citizen facing automated decisions.

Figure 1 illustrates the approach chosen to include human in the decision-making loop: the use of visual representations that encode large, complex datasets into interactive graphics. On the left, a parametric space of all the possible model configurations is given [Cancino12]. On the right, the possible outcomes of a model are shown as trajectories that can be chosen by a user as a good candidate.

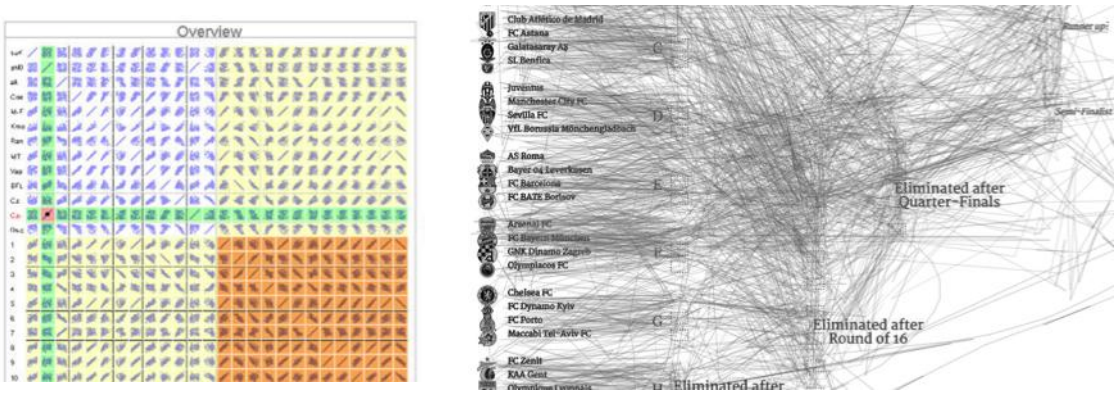

Figure 1. Left: visualization of parametric space for models configuration [Cancino12]. Right: interactive visualization of all models outcomes [Vuillemot16].

The application of the work will be on predictive models, which are a particular class of statistical models (both parametric and non-parametric) that give the probability of an outcome, after training on historical data. Application areas range from climate temperature predictions, financial time series, to image classification. This work is a continuation of previous efforts conducted at LIRIS on both Human Computer Interactions for sport data predictions [Vuillemot16] by analyzing mouse movements to choose models (Figure 1, right), as well as in using image classification models [Derrode16].

Research subject, work plan:

The scientific question we are investigating is ***how humans can be involved in the computational loop to improve models quality and acceptability, using interactive visualizations***. By improving, we mean finding better input parameters in the parameters research space, over current best practices or automated methods (e.g. parametric optimization). Models acceptability will be addressed by a process of white-boxing models, which can have important in case of post-mortem analysis of errors (like a plane crash), using interactive visualization as well.

The PhD work plan includes the following phases:

- Select a series of predictive models in various domains; identify their strengths and weaknesses, and perform a qualitative user study to understand their input parameters space, automated configurations opportunities, quantitative and qualitative evaluation of the output, and feedback loops for model controls.
- Design novel input parameters space visualization, with interactions to let human explores and visualize in real time the impact on the output. Heuristics and approximation functions will be required to guaranty a quality of result (<100ms).
- Implement the visualization and its underlying infrastructure (preferably using a web technology stack).
- Conduct testing, validation and evaluation with domain experts of the previous visualization. Iterate when needed to meet experts' satisfaction and a contribution over the state of the art tool (eventually participate to a benchmark or international competition such as the VAST Challenge).
- Design and Provide agonistic white-box visualizations to understand models execution, and to communicate models to a non-expert audience. Also provide other strategies for models that remain black boxes.

The candidate will benefit from working in a laboratory gathering over 300 researchers who are directly or indirectly working on computational models with big data. The laboratory also has a network of collaborators in this area that can be leveraged for international (Boston MA, Washington DC, Seattle, WA) and industrial collaboration. Publications are expected in to major conferences and journals in the filed (SIGCHI, VGTC, PVLDB, etc.). Coding and/or design skills are also a strong component of the PhD.

References:

[Pretorius11] Pretorius, A. Johannes, et al. "Visualization of parameter space for image analysis." IEEE Transactions on Visualization and Computer Graphics 17.12 (2011): 2402-2411.
[Cancino12] Cancino, Waldo, Nadia Boukhelifa, and Evelyne Lutton. "Evographdice: Interactive evolution for visual analytics." 2012 IEEE Congress on Evolutionary Computation. IEEE, 2012.
[Mühlbacher14] Mühlbacher, Thomas, et al. "Opening the black box: Strategies for increased user involvement in existing algorithm implementations." IEEE transactions on visualization and computer graphics 20.12 (2014): 1643-1652.
[Sedlmair14] Sedlmair, Michael, et al. "Visual parameter space analysis: A conceptual framework." IEEE transactions on visualization and computer graphics 20.12 (2014): 2161-2170.
[Ribeiro16] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why Should I Trust You?": Explaining the Predictions of Any Classifier." KDD (2016).
[Derrode16] Derrode, Stéphane, and Wojciech Pieczynski. "Unsupervised classification using hidden Markov chain with unknown noise copulas and margins." Signal Processing 128 (2016): 8-17.
[Vuillemot16] Vuillemot, Romain, and Charles Perin. "Sports Tournament Predictions Using Direct Manipulation." IEEE Computer Graphics and Applications (2016).