# Introduction

## Motivation

The work of this PhD is in the context of a CIFRE[1] thesis with Schneider Electric and two public research laboratories, the LIG[2] and the GIPSA-lab[3]. Within Schneider Electric, the PhD took place in the Analytics for Solutions (A4S) team, part of the Technology and Strategy entity whose main missions are à compléter par Sylvain . Among the wide activities of the A4S team, in the context of physical system modeling (*e.g.*, building, sensor network, Internet of Things [Naj+12]; [Ngu+12]; [YG08]), two topics are at least studied: modeling by physical models (white/grey box) and modeling by machine learning algorithms (black-box). With the increase of the amount of data and sensors that collects data, modeling accurately systems through *a priori* equations (white/grey box) for some prediction tasks has become more and more difficult. Within the vast amount of applications in Schneider Electric, some applications in particular involve temporal data, *e.g.*, forecasting the energy consumption in a building, virtual sensors, fault detection. More generally, Schneider Electric, like many other companies and other diverse application domains (medicine, marketing, meteorology, etc.) has taken a growing interest these last decades in machine learning problems (classification, regression, clustering) that involves time series of one or several dimensions, of different sampling, of two or more classes, etc. A time series can be seen in signal processing and in control theory as the response of a dynamic system. Contrary to static data, time series are more challenging in the sense that the temporal aspect (*i.e.*, order of appearance of the observations) is an additional key information.

A compléter par Sylvain

## Problem statement and contributions

In this work, we focus on classification problems of monovariate time series (1 dimension) with a fixed sampling rate and of same lengths. Among the wide variety of algorithms that exist in machine learning, some approaches (*e.g.*, $k$-nearest neighbors) classify samples using a concept of neighborhood based on the comparison between samples. In general, the concept of 'near' and 'far' between samples is expressed through a distance measure. Time series can be compared based not only on their amplitudes like static data but also on other characteristics or modalities such that their dynamic or frequency components. Many metrics for time series have been proposed in the literature such that the euclidean distance, the temporal correlation, the Fourier-based distance, etc. [BC94b]; [AT10b]; [SS12a]. A detailed review of the major metrics is proposed in [MV14]. In general, the existing metrics involve one modality at the

[1] Conventions Industrielles de Formation par la REcherche
[2] Laboratoire d'Informatique de Grenoble
[3] Grenoble Images Parole Signal Automatique

global scale (*i.e.*, implying systematically all the time series observations). We believe that the multi-scale aspect of time series, not present in static data, could enrich the definition of the existing metrics.

In this work, our objective is to learn a combined multi-modal and multi-scale time series metric for a robust $k$-NN classifier. The main contributions of the PhD are:

- The definition of a new space representation: the dissimilarity space which embeds a pair of time series into a vector described by basic temporal metrics.
- The definition of basic temporal metrics that involves one modality at one specific scale.
- The learning of a multi-modal and multi-scale temporal metric for a large margin $k$-NN classifier of univariate time series.
- The definition of the general problem of learning a combined metric as a metric learning problem using the dissimilarity representation.
- The proposition of a framework based on Support Vector Machine (SVM) and a linear and non-linear solution to define the combined metric that satisfies at least the properties of a dissimilarity measure.
- The comparison of the proposed approach with standard metrics on a large number of public datasets.
- The analysis of the proposed approach to extract the discriminative features that are involved in the definition of the learned combined metric.

## Organization of the manuscript

The first part makes a review of existing methods in machine learning and metrics for time series. The first chapter presents classical approaches in machine learning. In particular, we recall the general principle, framework and focus on two standard machine learning algorithms: the $k$-nearest neighbors ($k$-NN) and the Support Vector Machine (SVM) approach. In the second chapter, we review some basic terminology for time series and recall three types of metrics proposed at least for time series: amplitude-, behavior- and frequential-based. Then, we review the concept of metric learning for static data and focus on a framework of metric learning for nearest neighbors classification proposed by Weinberger & Saul [WS09b].

The second part of the manuscript propose a multi-modal and multi-scale metric learning ($\text{M}^2$TML) method. In the third chapter, we formalize the general optimization problem based on a new space representation, the dissimilarity space. We present a multi-modal and multi-scale time series description and their corresponding basic metrics. From the general formalization, we propose three different formalizations. The first and second proposition involve different regularizers, allowing to learn an *a priori* linear or non-linear form of the combined metric. The third proposition presents a framework based on SVM and a solution

to build the combined metric, in the linear and non-linear context, satisfying at least the properties of a dissimilarity measure. Finally, Chapter 5 presents the experiments conducted on a wide range of 30 public and challenging datasets, and discusses the results obtained.