

# INTRODUCTION

## 1.1 Biosignals and Challenges of a Data-Driven Society

In recent years, the continuous increase in accessible wearable technology has contributed to a significant amount of data available. The continuous production of data from wearable devices through the usage of mobile phones, smartwatches, hearables, wristbands, and other non-invasive wearable sensors has provided a valuable quantity of information. This data often comes in the form of time series, being one of the most common data types in nature [33]. As reported in *Tankovska et al.*, the wearable devices usage has more than doubled in the interval between 2016 and 2019, reaching 722 million [68], leading to a large volume of time series data being gathered in all possible scenarios, by monitoring patients in healthcare institutions [76, 67, 47, 11, 51, 8], tracking everyday activities of humans [14, 15, 4], recording machines in industrial processes or workers motion while performing their tasks [62, 61].

It has never been so easy to gather data about any aspect of our life, work, education, society, or industry. Of course, having relevant information about a subject is beneficial, but the overwhelming amount of data brings tremendous challenges in the ability to save, process, analyze and retrieve interpretable and meaningful information from which we can act upon [69]. Ultimately, it becomes even harder to have data well structured and labeled, considering that it is a sensitive and time consuming process, and complexity increases with data quantity. In the work of *Roh et al.* it is mentioned that data scientists only rely on a small portion of the available datasets because it is too expensive to label all the data available [59], and this is just an example of how much data can be unused. This is particularly problematic when developing machine learning applications, as data should be correctly pre-processed and labeled to be sure not to include noise, artifacts or mislabeled segments of the signal (Garbage-in Garbage-out - GIGO) [59].

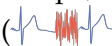
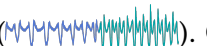

We believe that we should do more with the data we have and for that, tools should be available to support and help analysts to accelerate the process of information retrieval from time series. Both for analysts that are (not) familiar with time series data mining. Having more informative, expressive, and intuitive methods for the analysis of such data,

increases the speed of analysis for experienced analysts and promotes the democratization of this analysis for less or non-experienced analysts [42].





In this thesis, we propose novel methods that contribute to the information retrieval problem on time series. The reader will find two main domains of methods focused in (1) *unveiling the grammar of time series* and (2) *a language for time series data mining*. The proposed methods are designed to help in (1) explaining the story that originated the time series through a visual representation that highlights its structure and organization and (2) making the search for patterns and events with expressive queries, moving towards human interpretable and readable time series analysis. As the reader may notice, our topics merge the concepts of time series with text. Then, the reader may appreciate that we dwell on what we believe is the linguistic nature of time series.

To Discuss

## 1.2 Linguistic Nature of Time Series

Time Series are a visual domain, from which humans can create a good intuition. It is inherent to our ability to see relevant structures and patterns. The reader can imagine a recurrent shape, such as the QRS complex of an ECG signal that is interrupted by a noisy segment () . When interpreting this signal, we see that it has 3 representative segments and that the first is very similar to the third one. We could then represent the signal by A B A. Some shapes may be harder to distinguish, for instance, consider an accelerometer signal of a subject while walking and shifting to jogging regime () . Or a change in the shape of the arterial blood pressure (ABP) signal when there is a change in the subject's posture () . In both cases, the signal has 2 structures of a similar representative periodic pattern (A B).

pattern

This visual intuition is also very clear when a (non-)experience analyst is searching for specific shapes or patterns in time series. The reader may agree that scientists or other professionals often resort to describe the shape they are looking for. For instance, a physician may say "I am searching for the T-wave, that represents the large peak" () or "I am searching for the QRS complex, that looks like a sharp peak followed by a sharp valley" () . This visual intuition also happens when analysts are trying to find differences between classes of signals. For instance, the following shapes (1)  and (2)  are different because "shape 1 has a peak where shape 2 doesn't".

Time series are carriers of information and the presence of a change in the regimes of a time series or the presence of a specific shape in a segment of a time series may be associated with a specific occurrence in the physical world and be attributed to a meaning. This notion of structure and meaning is a good approximation of what represents the foundation of a language: grammar and meaning [16].

Grammar is generally defined as the book of rules that constitutes the structure of a language and is modeled by the morphology and syntax [16]. The first is the structure of words, how these are built or morphed from individual symbols, while the latter consists in

organizing words in sequences to form larger linguistic units, such as sentences. Just as a language has morphological and syntax rules that represent its structural information, time series <sup>can be</sup> are also organized by a formal structure of ordered subsegments with specific morphological characteristics, organized to build larger segments. Our first topic is related to *unveiling* this structure with methods that can parse it.

In addition to a *grammar*, a language also has *meaning*. The *meaning* on time series depends on the context and what occurred in the physical world that is seen on the signal. Specific occurrences might be attributed a specific meaning by an analyst, as we have seen above with the physician example. In this work, we explore a language to *translate* time series into text and use this textual information as an expressive way of searching for meaningful events and patterns. This is related to our second topic, which focuses on using language in time series data mining tasks.

Until now, we have been using the term *time series*, but the thesis is entitled *A Language for Biosignals*. Having explained how time series have a *linguistic nature*, we now focus our attention to a specific domain of time series, *biosignals*, which are time series that come from the *human body*, such as the *heart* (ECG - ), *muscles* (Electromyogram (EMG) - ), *brain* (Electroencephalogram (EEG) - ) or even movements (Inertial Motion Unit (IMU) - ). Considering that the tools developed can be employed in any time series domain, we will use this term instead, but we will give most of our examples from occupational *biosignals*, with a special interest in showing how these can be helpful and meaningful in the context of occupational health.

To discuss

### 1.3 Biosignals: Context and Relevance in Occupational Health

This thesis was developed in a ~~strong~~ partnership with the *Ergonomy* team from Volkswagen Autoeuropa. Therefore, the central domain of the application regarded the analysis of *biosignals* from workers to retrieve meaningful information about their occupational risk status and prevent Work Related Musculoskeletal Disorder (WMSD)s. WMSDs prevail as the most common occupational disease in the European Union. These have a global impact on the well-being of individuals and their quality of life in a range of working sectors [34], accounting for the second-largest responsibility to disability worldwide [45]. These are especially prevalent as upper limb or neck disorder (with 42% of all WMSD cases reported) [28] in several industry sectors, such as textile and automotive, where production processes with pre-defined motions and actions have a repetitive/cyclic nature. This has a negative impact on the risk to develop musculoskeletal disorders, with tremendous consequences to both workers and companies, leading to absenteeism, early retirement, and loss of productivity [70, 71].

Several strategies have been implemented to identify, regulate and prevent occupational risks in manufacturing industries, such as (1) the inclusion of job rotation schedules, which promote a variation of the exposure throughout the working day [5, 58] and (2) screening tools, for the assessment of occupational risk exposure, e.g. Occupational Repetitive

Action (OCRA), Rapid Upper Limb Assessment (RULA) or the Ergonomic Repetitive Worksheet (EAWS) [54, 49, 63]. Nevertheless, these strategies are not optimal because they (1) are not automated, relying on observational methods and dedicated personnel to inspect video records; (2) are not objective measures; (3) do not take into account differences among the worker's population, as anthropometric, age and experience variability; and (4) present single scores, being insufficient to explain the factors that contributed to this risk. With the advent of Industry 4.0, more companies are using modern strategies that follow digital solutions to provide direct and objective quantitative measures [60]. An example of these incentives is the usage of *biosignals*, with wearable inertial devices for physiological, motion, and posture tracking of workers.

From *IMU*, time series can be collected and relevant information can be directly measured, e.g. position and velocity of each body segment, postural angles between joints, and gait parameters, making these important for ergonomics studies [13, 72]. There are some limitations to using *IMU*, mostly related to the long-term bias (sensor drifting) arising from long acquisitions and the empirical process to fine-tune sensor fusion techniques. Other systems can be used for motion capture, such as camera-based methods, but these rely on a fixed setup of cameras, which is unmanageable in real industrial scenarios [61]. In addition to motion sensors, the inclusion of physiological sensors, such as *ECG*, *EMG* and even *functional Near InfraRed Spectroscopy (fNIRS)* can give reliable evidence of other occupational health variables, namely cardiovascular load, muscular activation, cognitive effort and fatigue [66, 41, 32, 23].

The usage of biosignals in this context can play an important role in supporting the decision of ergonomists and other professionals in the industry. To develop systems that can use physiological, motion, and postural data for direct risk assessment and reporting, several challenges arise in the time series data mining domain. For instance, considering the periodic nature of most manufacturing tasks, risk factors are calculated by working cycle. Therefore, methods should be developed to identify working cycles with some variability in their periodicity. In addition, real occupational scenarios might have interruptions or changes in the working behavior, due to abrupt production stoppage, shift breaks, or even because the worker shifted to another workspace that has a different movement pattern.

Other questions also arise by ergonomists, such as *can we find a pattern that has a sharp rise in the IMU from the arm?* or *when the worker is using a hand tool to rotate a screw, can we see a periodic pattern on the IMU from the hand?*, which represent specific patterns with a descriptive shape that can be seen on the signals and are specific of a task. These events can be relevant to studying their precise impact on the worker's occupational exposure. Having ways to detect these patterns is of great relevance as well. In this study, we will show how the proposed solutions can have an impact on these problems, and how they contribute to providing relevant visual feedback for information retrieval from the occupational data and make the search for specific patterns more intuitive and expressive, even for non-experienced data analysts, such as ergonomists.

## 1.4 Research Paths

The previous sections introduced the topics explored in this thesis for information retrieval on time series, our main motivations to develop the proposed methods, and how these can have significant contributions in the biosignals domain, more specifically for occupational health data.

The work in this thesis contributed to all layers related to time series, from the moment data is acquired (*sensing*), processed for information retrieval (*analysis*), and how it is used to act upon (*decision making*). From these, the presented work in this document will especially address the development of methods for information retrieval (*analysis*) from time series for better *decision making*.

1. **Sensing** - Explore in depth the available technology to measure motion and postural variables in occupational scenarios for risk assessment. This will take into account which variables are associated with a risk, based on ergonomic standards. These measures are returned as time series, which are processed in the topic *analysis*;
2. **Analysis** - In this topic, three main research paths are explored with specific research topics. **A** - (1) study how to perform structural information retrieval in time series for segmentation based on change points and periodic points and (2) how are the segments related based on their similarity. For this, we applied a feature-based transformation of the time series and similarity-based measures to make a meaningful visual representation, from which the segmentation points can be extracted and the relationship between segments can be made. **B** - explore a symbolic representation of time series and a word feature-based representation of time series, studying how these can be used for more expressive and intuitive pattern search with the help of regular expressions and ultimately natural language. **C** - From the textual representation of time series, study if we can make a higher leveled distance measure, following standard text mining methods. The resulting outputs of these methods can ultimately be used to be more aware of why a signal is different from the others.
3. **Decision Making** - Discuss how the developed methods can contribute to more aware and informed decisions. Considering the outputs of the methods developed in research path A, how can the analyst gain intuition over the structure of the data associating it with what happened in the physical world. In what regards to research path B, how expressive is the process of searching for specific events and patterns with the proposed linguistic-based search methods.

## 1.5 Thesis Structure

This thesis provides a detailed description and explanation of the research work developed during the Ph.D. program. It is organized into nine Chapters, each contributing to telling

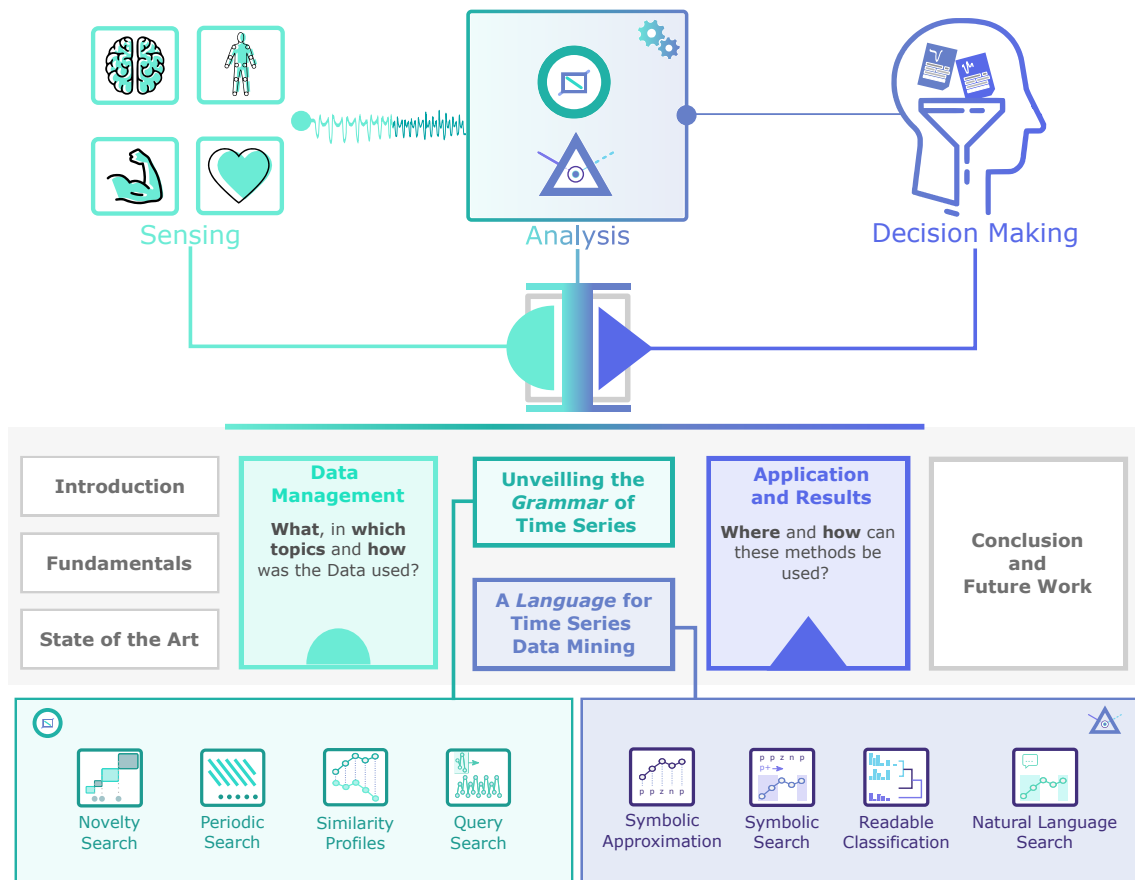


Figure 1.1: General topics contemplated on this thesis and structure of the document. It highlights the 3 layers of involvement related to time series: Sensing, Analysis, and Decision Making, focusing on the Analysis layer, which includes two major topics subdivided into 4 sections each.

the story of this thesis. The reader may appreciate Figure 1.1, which illustrates a guideline of the structure of this work, with a short description of each Chapter's topics and content. **Chapter 1** introduced the main motivations, goals, and context for the development of the proposed methods. **Chapter 2** provides the reader with the fundamental definitions and knowledge needed to have a clear picture of what is developed in this work. **Chapter 3** depicts the state-of-the-art works related to what we developed, namely in the topics of segmentation, summarization, pattern/event search and classification. **Chapter 4** describes the data we used, explaining its source for both private data and publicly available data, for which purposes it was used and how it was used in this work. **Chapter 5** explains the algorithm developed for time series structural information retrieval (Unveiling the *Grammar* of Time Series) and provides examples of its usage for novelty segmentation, periodic segmentation, similarity profiles, and query search. *Chapter 6* covers the usage of language for time series data mining (*A Language* for time series data mining), more specifically introducing a novel symbolic approximation and how it can be used for pattern search and classification. In addition, it also explains how to use natural language with a word-feature-based representation of time series. **Chapter 7** shows the application of

the previous methods to an exhaustive set of examples, namely from the occupational scenario, and presents major results. In addition, this chapter also provides a general discussion of these <sup>examples</sup> and how the proposed methods can be used for the benefit of the analyst. **Chapter 8** gives an overall remark on the outcomes of this thesis and a reflection on the contributions that the developed methods have in making time series preparation and data mining more expressive, quicker, and more practical for an ever-increasing number of data available. It also provides ~~the reader with~~ a clear idea of which are the future paths for this work in terms of novel applications and performance improvement.