Abstract: This tutorial gives an overview of data mining techniques suitable for application to sensor and measurement data. Time series and sequence analysis are particularly relevant to mining sensor data. The emergence of cyber physical systems has brought about large volumes of sensor data being streamed in real-time. Characteristic for such data sets is: that they are temporally ordered, fast changing, massive, they are fundamentally parallel in nature, potentially infinitely long and we need to ensure the physical meaning of the knowledge which is extracted. Mining such data delivers a more complete perspective of complex systems. There main issues addressed are:

1. Physical meaning: Linear differential operators are used to model physical systems. It is shown that forward problems can be solved by applying an FIR filter and inverse problems with an IIR structure. In this manner the sensor data can be processed in real time while ensuring that the solutions fulfill the physical models. Virtually all present data mining techniques use statistical data models based on correlation as a measure of significance when discovering knowledge; however, in a physical system where a sensor is making an indirect measurement the physical model is a prerequisite if meaning and relevance is to be associated with causality and not just correlation.

2. The concept: The concept of lexical analysis for sensor data is introduced; this is an extension of the symbolic aggregation approximation method. The processed sensor data is approximated by a set of symbols, then tokens with predicates are extracted from the symbol-time-series, yielding a sequence which is a compressed symbolic approximation of the sensor data. The symbolic data is easily combined with meta-data enabling symbolic queries about the state of the complex system. Additionally, the lexical analysis implements dynamic time warping (DTW), so that sequence matching can be performed with a desired tolerance in the time axis distortion.

3. The implementation: The concepts for both software and systolic-slice implementations are presented. In particular, the systolic-slice can be targeted directly to an FPGA yielding a real-time parallel implementation of the lexical analysis, e.g., with a typical FPG approximately 250 sensor channels can be processed in parallel with a sampling frequency of 100 kHz. The resulting symbolic model is suitable for querying and OLAP type investigation.

4. The philosophical background: The relationship of the lexical analysis to the development of natural speech, as understood by the philosophy of phenomenology, is introduced. The philosophical basis is that human speech evolved to describe our experience of phenomena from sensory experience. The western phenomenology founded by Edmund Husserl (1859 – 1938), a physicist and extended by Martin Heidegger (1889 –1976) proposed that we obtain all knowledge via interaction with objects. The eastern phenomenology, Vasubhandu (~ 400), on the other hand shows that we are never in direct contact with an object, but are always in contact with a mental model for the object. In this case modeling of the object being observed is central to knowledge acquisition.

The data mining approaches are demonstrated with the remote monitoring of heavy plant and machinery. The methods enable the determination of the state of the machinery and the automatic identification of operations. This information is used to implement reactive control of the plant and to optimize performance.