SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Healthcare Data Analytics

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

Somaiya TRUST

- Sensors measure physical attributes of the world and produce signals, i.e., time series consisting of ordered measurements of the form (timestamps, data elements).
- For example, in intensive care, respiration rates are estimated from measurements of the chest impedance of the patient.
- The resulting time series signals are consumed either by a human or by other sensors and computing systems.
- For instance, the output of the chest impedance sensor can be consumed by an apnea detection system to produce a signal measuring apnea episodes.
- The data elements produced by sensors range from simple scalar (numerical or categorical) values, to complex data structures.
- Examples of simple data elements include measures such as hourly average of temperature in a given geographical location, output by a temperature sensor.
- Examples of more complex data elements include summaries of vital signs and alerts measured by a patient monitor sensor in a medical institution.

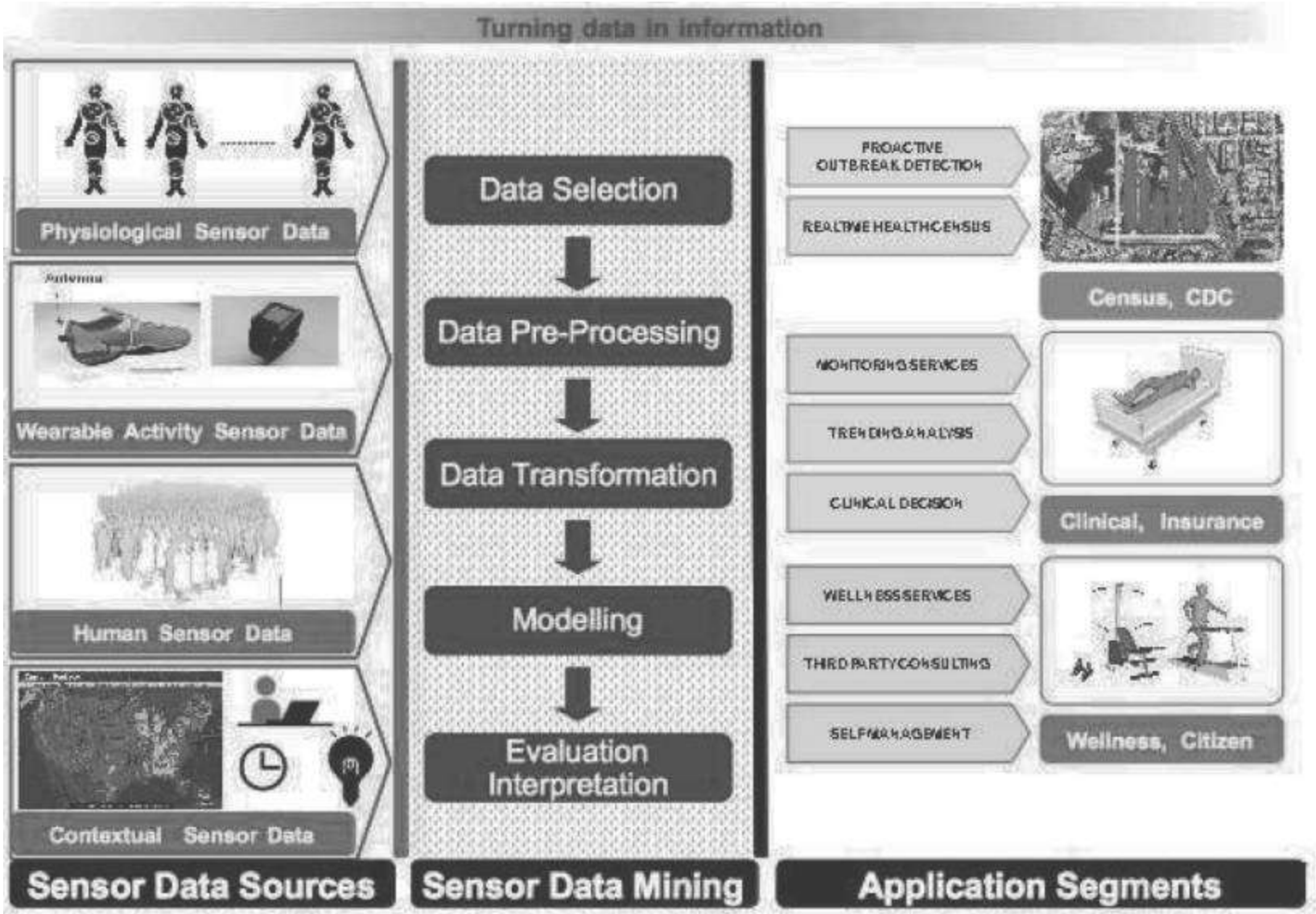
Taxonomy of Sensors Used in Medical Informatics

- *Physiological sensors*: These sensors measure patient vital signs or physiological statistics.
- They were first used to measure vitals on astronauts before appearing in medical institutions, at the bedside in the 1960s.
- Today, physiological sensors are also available outside medical institutions, even on pervasive devices (e.g., iPhone heart rate monitor applications that make use of smartphone cameras

- *Wearable activity sensors*: These sensors measure attributes of gross user activity, different from narrowly focused vital sign sensors.
- Good examples are accelerometers used for gait monitoring. Shoe manufacturers like Nike have enabled many of their running shoes with sensors capable of tracking walking or jogging activities.
- Most smartphones are also equipped with accelerometers and several wellness management applications leverage these sensors.

- *Human sensors*: Humans play an integral role in the sensing process. For instance, physicians introduce important events that relate to the patient health status during examinations.
- Lab technicians follow rigorous processes to provide blood content information. Self-reporting (i.e., patients monitoring their health parameters) is also used in the management of chronic illnesses like diabetes.
- More recently, with the emergence of social media and pervasive computing, people use mechanisms like Web searches and Twitter to generate reports on important health-related events.

The sensor data mining process.



- *Contextual sensors*:
  - These sensors are embedded in the environment around the user to measure different contextual properties.
  - Examples include motion detection sensors, audio and video sensors, temperature sensors, weather sensors, etc.

Challenges in Mining Medical Informatics Sensor Data

- As with standard data mining procedures, healthcare mining is typically performed in five stages:

- 1. Data Acquisition: This includes operations involved in collecting data from external sensor data sources.
- 2. Data Preprocessing: This includes operations applied to the data to prepare it for further analysis. Typical preprocessing operations include data cleaning to filter out noisy data elements, data interpolation to cope with missing values, data normalization to cope with heterogeneous sources, temporal alignment, and data formatting.
- 3. Data Transformation: The includes operations for representing the data appropriately and selecting specific features from this representation. This stage is often called feature extraction and selection.

- 4. Modeling: This stage, also called mining applies knowledge discovery algorithms to identify patterns in the data. Modeling problems can be classified into six broad categories: (1) anomaly detection to identify statistically deviant data, (2) association rules to find dependencies and correlations in the data, (3) clustering models to group data elements according to various notions of similarity, (4) classification models to group data elements into predefined classes, (5) regression models to fit mathematical functions to data, and (6) summarization
models to summarize or compress data into interesting pieces of information.

- 5. *Evaluation*: This stage includes operations for evaluation and interpretation of the results of
the modeling process.

Sensor Data Mining Analytical Challenges at Each Stage of the Data Mining Process

(I) Acquisition	*(II) Pre-processing*
lack of data standards	data formatting
lack of data protocols	data normalization
data privacy	data synchronization
(III) Transformation	*(IV) Modeling*
physiological feature extraction	sequential mining
feature time scales	distributed mining
unstructured data	privacy preserving modeling
	obtaining ground truth
	exploration-exploitation trade-offs
(V) Evaluation and Interpretation	
Model expressiveness	
Process and data provenance	

Challenges in Healthcare Data Analysis

- Despite several standardization efforts, medical sensor manufacturers tend to design proprietary data models and protocols to externalize sensed signals.
- In healthcare, standard bodies like HL7 and the Continua Health Alliance address data modeling issues while several IEEE standard protocols address device interoperability issues.
- However, there is a lack of incentives for sensor data manufacturers to adhere to these standards. With this lack of adherence to standards, mining medical sensor data across multiple data sources involves several nontrivial engineering challenges, and the design of custom solutions specific to each sensor data mining application.
- Another key challenge in the acquisition process is related to the protection of user privacy. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) defines regulations on access to health data.
- By law, data mining applications that leverage this data must comply with these regulations. Data de-identification and de-anonymization techniques are often required to comply with HIPAA.
- Privacy preserving data mining techniques may also be used to extract information from sensor data while preserving the anonymity of the data.

- Acquisition Challenges In a clinical setting such as the ICU, includes different types of physiological sensors (e.g., ECG sensors, SpO2, Temperature sensors), contextual sensors (e.g., RFID sensors linked with care providers, video and cameras) and human sensors (e.g., care-provider notes, entries in the Electronic Medical Records).
- More recently, with the emergence of wearable devices, and network connectivity, additional information is provided (even in nonclinical settings) by activity sensors (e.g., wearable devices such as cell phones) and completely nontraditional sources of information (e.g., community discussions in healthcare-related sites, aggregated views of user searches, etc.).

- Acquiring and integrating this data is nontrivial because of the inherent heterogeneity and lack of standards and protocols. Physiological sensor manufacturers have mostly designed proprietary data models and protocols to externalize sensed signals, despite the efforts of standard bodies like HL7 and the Continua Health Alliance to address data modeling issues, and IEEE standard protocols to address device interoperability issues. Additionally, there is little standardization or interoperability studies of contextual and activity sensors, and data from healthcare providers is captured poorly, often requiring manual entry and transcription – all making the data acquisition task extremely complex.
- This has led to the emergence of data aggregators in ICU and EHRs for general clinical settings, however these aggregator solutions operate only on a narrow set of sources, and often do not interoperate with each other.
- Hence, mining medical sensor data across multiple data sources has involved several nontrivial engineering challenges, and the design of custom solutions specific to each sensor data mining application.

- These acquisition challenges are compounded by the need to provide privacy protection for this often very sensitive personal information.
- This includes conforming with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) act and providing appropriate controls with mechanisms for authentication, authorization, anonymization, and data de-identification.
- This also requires the design of privacy preserving data mining and analysis techniques.
- There are also several open, unresolved questions related to the privacy protection of data generated using nontraditional, contextual, and activity sensors.

Preprocessing Challenges

- Data in the real world is inherently noisy. The preprocessing stage needs to address this problem with sophisticated data filtering, sampling, interpolation, and summarization techniques (such as sketches, descriptive statistics to minimize the effects of noise.
- The preprocessing also needs to account for the heterogeneity of data, and the lack of standards adoption by medical sensor manufacturers. Indeed, data generated in different formats needs to be syntactically aligned and synchronized before any analysis can take place.
- Sensors report data with timestamps based on their internal clocks. Given that clocks across sensors are often not synchronized, aligning the data across sensors can be quite challenging. In addition, sensors may report data at different rates.
- For instance while the ECG signal is generated at several 100s of Hz, the EMR may only be updated hourly. Aligning these datasets requires a careful design of strategies. These preprocessing techniques need to handle different types of structured data such as transactions, numeric measurements, and completely unstructured data such as text and images, often jointly.
- It is critical that the preprocessing of these sources retains the appropriate correlation structures across sources, so that meaningful and subtle indicators of patient health can be detected.

- Furthermore, a semantic normalization is often required to cope with differences in the sensing process.
- As an illustration, a daily reported heart rate measure may correspond to a daily average heart rate in some cases, while in other cases it may represent a heart rate average measured every morning when the subject wakes up.
- Comparing these values in a data mining application can yield incorrect conclusions, especially if they are not semantically distinguished. All of these issues make the preprocessing task very complex

Transformation Challenges

- Data transformation involves taking the normalized and cleaned input data and converting it to a representation such that attributes or features relevant to the mining process can be extracted.
- This may include applying different types of linear (e.g., Fourier Transform, Wavelet Transform) and nonlinear transformations to numeric data, converting unstructured data such as text and images into numeric representations (e.g., using a bag of words representations, or extracting color, shape, and texture properties), and applying dimensionality reduction and de-correlation techniques (e.g., Principal Component Analysis), and finally summarizing the result with a set of representative features that can then be used for analysis and modeling.
- The choice of the appropriate transformations and representations for the features is heavily dependent on the task that needs to be performed.
- For instance a different set of features may be required for an anomaly detection task, as opposed to a clustering or classification task.
- Additionally, the choice of appropriate features requires understanding of the healthcare problem at hand (e.g., the underlying physiology of the patient) and often requires inputs from domain experts.
- For instance, in neurological intensive care environments, spectral decomposition techniques for feature extraction have been defined, in conjunction with domain experts, to aid the interpretation of electroencephalograms (EEG) signals for brain activity monitoring and diagnosis of conditions such as seizures

- In addition to such signals, human sensing adds different types of unstructured data that need to be effectively integrated.
- This includes textual reports from examinations (by physicians or nurses) that need to be transformed into relevant features, and aligned with the rest of the physiological measurements.
- These inputs are important to the data mining process as they provide expert data, personalized to the patients. However, these inputs can be biased by physician experiences, or other diagnosis and prognosis techniques they use.
- Capturing some of these aspects during the mining process is extremely challenging.
- Finally, there is a lot of external domain knowledge in open source repositories, medical journals, and patient guidelines that can be relevant to patient care, and features should be placed in context of this knowledge for appropriate interpretation.

Modeling Challenges

- There are several challenges that need to be overcome in the modeling stage of the data mining process for medical sensor data. First of all, the time series nature of the data often requires the application of sequential mining algorithms that are often more complex than conventional machine learning techniques (e.g., standard supervised and unsupervised learning approaches).
- Nonstationarities in time series data necessitate the use of modeling techniques that can capture the dynamic nature of the state of the underlying processes that generate the data. Known techniques for such problems, including discrete state estimation approaches (e.g., dynamic Bayesian networks and hidden Markov models) and continuous state estimation approaches (e.g., Kalman filters or recurrent neural networks) have been used only in limited settings.
- Another challenge arises due to the inherent distributed nature of these applications. In many cases, communication and computational costs, as well as sharing restrictions for patient privacy prevent the aggregation of the data in a central repository.
- As a result, the modeling stage needs to use complex distributed mining algorithms. In remote settings, there is limited control on the data acquisition at the sensor. Sensors may be disconnected for privacy reasons or for resource management reasons (e.g., power constraints), thereby affecting the data available for analysis.
- Modeling in these conditions may also require the distribution of analytic approaches between the central repository and the sensors. Optimizing the modeling process becomes a challenging distributed data mining problem that has received only limited attention in the data mining community.

- Modeling in healthcare mining is also hindered by the ability to obtain ground truth on the data.
- Labels are often imprecise and noisy in the medical setting. For instance, a supervised learning approach for the early detection of a chronic disease requires well-labeled training data. However, domain experts do not always know exactly when a disease has started to manifest itself in a body, and can only approximate this time. Additionally, there are instances of misdiagnosis that can lead to incorrect or noisy labels that can degrade the quality of any predictive models.
- In clinical settings, physicians do not have the luxury of being able to try different treatment options on their patients for exploration purposes. As a result, historical data sets used in the mining process tend to be quite sparse and include natural biases driven by the way care was delivered to the patient.
- Standard approaches are not well-equipped to cope with this bias in the data, especially as it is hard to quantify precisely. Furthermore, most studies in medical informatics are retrospective.
- Well-done prospective studies are hard to do, and are often done on small populations, limiting the statistical significance of any derived results.

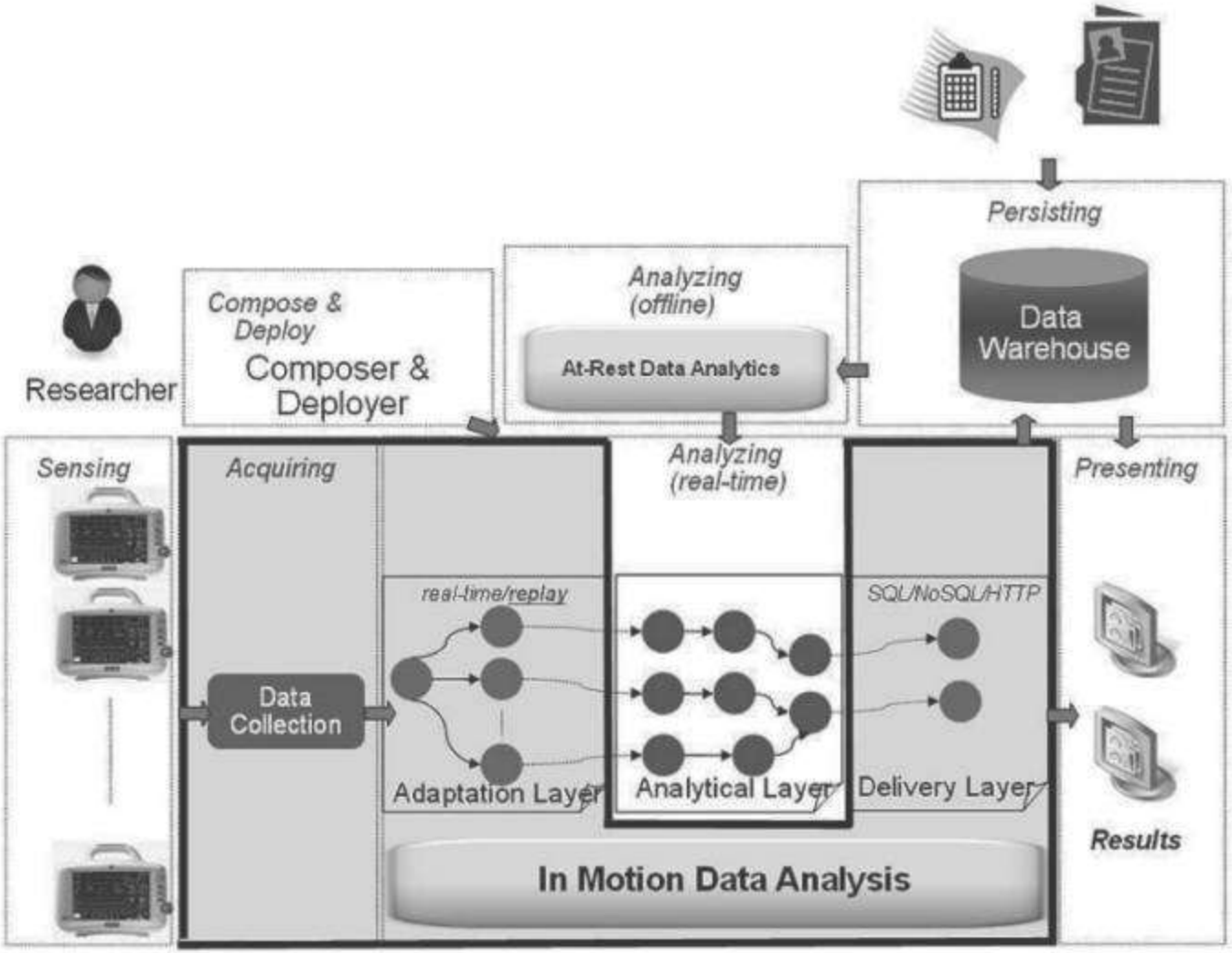
- **Evaluation and Interpretation Challenges**
- Data mining results consist of models and predictions that need to be interpreted by domain experts.
- Many modeling techniques produce models that are not easily interpretable.
- For example, the weights of a neural network may be difficult to grasp for a domain expert. But for such a model to be adopted for clinical use, it needs to be validated with existing medical knowledge.
- It becomes imperative to track provenance metadata describing the process used to derive any results from data mining to help domain expert interpret these results.
- Furthermore, the provenance of the data sets, and analysis decisions used during the modeling are also required by the experts to evaluate the validity of the results.
- This imposes several additional requirements on the selected models and analysis.

- **Generic Systems Challenges**
- Beyond analytical challenges, sensor data mining also comes with a set of systems challenges
  that apply to medical informatics applications.
- The mining of sensor data typically requires more than conventional data management (database or data warehousing) technologies for the following reasons:
  - The temporal aspect of the data produced by sensors sometimes generate large amounts of
    data that can overwhelm a relational database system.
  - For example, a large population monitoring solution requiring the real-time analysis of physiological readings, activity sensor readings and social media interactions, cannot be supported with relational database technologies alone.
  - Sensor mining applications often have real-time requirements. A conventional store-thenanalyze paradigm leveraging relational database technologies may not be appropriate for such time-sensitive applications.
  - The unstructured nature of some of the data produced by sensors coupled with the real-time
    requirements imposes requirements on the programming and analysis models used by developers of sensor data mining applications.

- Hence, sensor mining in healthcare requires the use of emerging stream processing system technology in conjunction with database and data warehousing technologies.
- Stream processing systems are designed to cope with large amounts of real-time data, and their programming models are geared towards the analysis of structured and unstructured sensor data. They are also time sensitive and analyze data within small latency bounds.

- Figure 4.2 presents an extended architecture for sensor data mining that illustrates this integration.
- The rationale behind this architecture is to use a stream processing system for the real-time analysis of sensor data, including the preprocessing and transformation stages of the analytical data mining process.

A generic architecture for sensor data mining systems.



- The sensor data acquisition is performed by a layer of software that interfaces with sensors and feeds into the stream processing system.
- The results of the transformation stage may be persisted in a data warehouse for offline modeling with machine learning techniques.
- The resulting models may be interpreted by analysts and redeployed on the stream processing platform for real-time scoring.
- In some cases, online learning algorithms may be implemented on the stream processing system.
- This integration of stream processing with data warehousing technologies creates a powerful architecture that addresses the system challenges outlined above.

Sensor Data Mining Applications

- **Intensive Care Data Mining**
  In 2003, it has been reported that intensivists have to handle over 200 variables, some of them being temporal, on a per patient basis to provide care. Anecdotal evidence tells us that this number has increased significantly since 2003 with the emergence of more and more sensing devices in critical care.
- Today, critically ill patients are often attached to large numbers of body sensors connected to sophisticated monitoring devices producing these large volumes of physiological data.
- These data streams originate from medical devices that include electrocardiogram, pulse oximetry, electroencephalogram, and ventilators, resulting in several kilobits of data each second. While these monitoring systems aim at improving situational awareness to provide better patient care with increased staff productivity, they clearly have introduced a data explosion problem.
- In fact, the vast majority of data collected by these monitoring systems in Intensive Care Units (ICUs) is transient.

- In talking with medical professionals, we learned that the typical practice in ICUs is for a nurse to eyeball representative readings and record summaries of these readings in the patient record once every 30–60 minutes.
- The rest of the data remains on the device for 72-96 hours (depending on the device's memory capacity) before it times out and is lost forever.
- Hospitals are simply not equipped with the right tools to cope with most of the data collected on their patients, prompting many to state that medical institutions are data rich but information poor.
- The potential of data mining in this area has been recognized by many. Several efforts are underway to develop systems and analytics able for the modeling of patient states and the early detection of complications.
- In general, early detection of complications can lead to earlier interventions or prophylactic strategies to improve patient outcomes.
- Early detection rests on the ability to extract subtle yet clinically meaningful correlations that are often buried within several multimodal data streams and static patient information, spanning long periods of time.

- Systems for Data Mining in Intensive Care Modern patient monitors have evolved into complex system that not only measure physiological signals but also produce alerts when the physiological state of the patient appears to be out of range.
- State-of-the-art patient monitors allow physicians to program thresholds defining normality ranges for physiological systems. For example, one can program a patient monitor to produce an audible alert if the oxygen saturation level of the blood is below 85%. The values of these thresholds are typically obtained from general guidelines or from data mining processes. Such simple alerting schemes are well known to produce very large numbers of false alarms.
- In, it is reported that more than 92% of alarms generated in an ICU are of no consequence.
- Furthermore, there are many complex physiological patterns of interest to physicians that cannot be represented by a set of thresholds on sensor data streams. Several research initiatives are addressing this problem with the design of platforms facilitating analysis beyond the simple thresholding capabilities of existing patient monitoring systems.

- One example is BioStream a system that performs real-time processing and analysis of physiological streams on a general purpose streaming infrastructure.
- The authors use ECG data along with temperature, oxygen saturation, blood pressure, and glucose levels as inputs into patient- specific analytic applications.
- The system supports a different processing graph (for analysis) per patient, where the graph can be composed of system supplied operators (functions) and user implemented operators.

State-of-the-Art Analytics for Intensive Care Sensor Data Mining

- State-of-the-art analytics and mining approaches for in-hospital sensor data monitoring (Figure 4.3) tend to generate innovations on data preprocessing and transformation.
- Modeling is typically done with well-known families of machine learning techniques such as classification, clustering, and dynamic system modeling with sequential learning.
- These analytical techniques often attempt to derive features from physiological time series to model the *inflammatory response* of the body, as it is known to be highly correlated with early sign of complications in general.
- The inflammatory response is a reaction from the body to different harmful stimuli such as pathogens, various irritants, or even damaged cells.
- Hence, accurate modeling of it enables a wide range of early detection applications in intensive care. In particular, devastating complications such as sepsis are known to produce an inflammatory response well before the appearance of clinical symptoms

- **Sensor Data Mining in Operating Rooms**
  Data mining applications that relate to operating rooms tend to focus on the analysis of Electronic Medical Record data where most sensor data inputs are filtered and summarized.
- For example, in, EMR data is used to improve the efficiency of operating rooms, in terms of scheduling (start times, turnover times) and utilization.
- In, knowledge management and data mining techniques are used to improve orthopedic operating room processes, yielding more effective decision making. A few researchers have reported applications directly mining physiological sensor data produced by operating room monitoring systems.
- Exceptions are presented in where the authors correlate EEG signals with cerebral blood flow measurements for patients undergoing carotid endarterectomy.
- This finding is quite valuable as it proves that EEG signals can be used to monitor complex mechanisms including cerebral blood flow for this patient population.
- In, machine learning techniques are proposed for the closed-loop control of anesthesia procedures.
- In, the authors present a prototype of a context-aware system able to analyze patient data streams collected in an operating room during surgical procedures, to detect medically significant events, and apply them in specific EMR systems.

- **General Mining of Clinical Sensor Data**
- Recent stimuli from the federal government and the increased ease of adoption of electronic health records system has led to widespread use of EHR in clinical practice.
- Large providers such as EPIC and McKesson have essentially unified the elements of data entry by having common platforms, although the use of free text and contextual rather than templated data is more common.
- EHR systems are a unique healthcare sensor, since real-time data and vast data troves are similar to other sensors, yet the relatively unstructured data makes it difficult to view this as a typical sensor.
- They typically contained structured and unstructured comprising of all the key administrative clinical data relevant to patients, demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, diverse test results, and radiology reports.

- Furthermore, there are no widely accepted standards for the representation of all these data points stored in EHR systems. Several code systems (e.g., ICD-9, ICD-10, CPT-4, SNOWMED-CT) and interoperability standards (e.g., HL7, HIE) are in use by many systems but there are no overarching standards that EHR vendors are adhering to.
- Despite this lack of global standardization that is hindering the realization of very large-scale data mining, many researchers are spending considerable efforts to analyze these data sets to improve healthcare in general.
- Mining of such sensors have been undertaken by various groups. The use of EHR mining todetect delays in cancer diagnosis, hospital-acquired complications, and the ability for groups to develop high throughput phenotyping to identify patient cohorts based on modular and consistent resources has been reported

Question ?