MONITOROLOGY – THE ART OF OBSERVING THE WORLD

MIROSLAW MALEK

Advanced Learning and Research Institute (ALaRI), Faculty of Informatics, Università della Svizzera italiana, via Buffi 13, 6904 Lugano, Switzerland

(received: 21 June 2021; revised: 11 August 2021; accepted: 13 August 2021; published online: 30 October 2021)

Abstract: We focus on the art of observing the world by electronic devices such as sensors and meters that, in general, we call monitors. We also define main monitoring objectives and pose five challenges for effective and efficient monitoring that still need a lot of research.

In the era where compute power like electricity is easily available and easy to use across the globe, and big data is generated in enormous amounts at ever-increasing rates, the question, what to monitor and how, will become ever more relevant to save the world from flood of meaningless, dumb data, leading frequently to false conclusions and wrong decisions whose impact may range from a minor inconvenience to loss of lives and major disasters.

Keywords: Data Acquisition, Data Analytics, Monitoring, Big Data, Internet of Things

DOI: https://doi.org/10.34808/tq2021/25.2/a

Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?

1. Introduction

Observation, monitoring or data acquisition, followed by data collection and analysis, has been the most common and perhaps the most successful scientific method since the beginning of times. It is not only used in science to observe the world and state hypothesis but it has been used virtually in all domains and all walks of life from archeology and business to physics and zoology. Formally, monitoring is observation and collection of relevant data about the current state of a system under study. The purpose of monitoring may vary from noble causes such as better understanding of the world and saving lives to dictatorship, espionage and blackmail including impinging on privacy. Monitoring is also used to ease or optimize control of machines including robots and vehicles. In electronic and mechanical systems (hardware) typically physical features are measured such as temperature, load or pressure by sensors while in software log files or probes are used. In this article, we focus mainly on technical aspects of monitoring but the ethical issues are equally vast and diverse and therefore require a separate treatment.

In the flood of data generated daily, it is not easy to filter out the relevant information, but even more challenging is to infer the knowledge, not mentioning the wisdom that even teams of experts are not able to derive in majority of real-life situations (see Figure 1). Although several classification methods ranging from statistics and machine learning to pattern recognition and data mining exist, knowing what to collect regarding data or information might be more effective than a particular method itself.

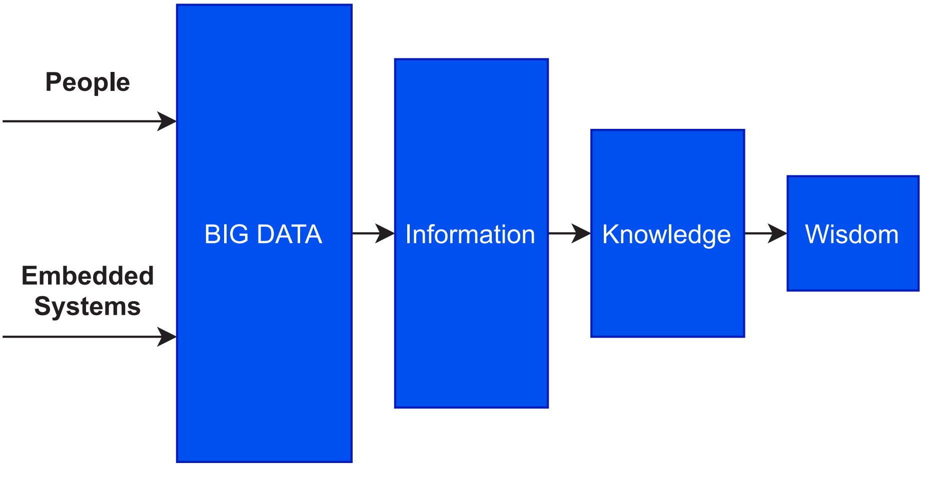

Figure 1. Challenge of getting Small Data out of Big Data in a form of information, knowledge or wisdom

The biggest generators of the masses of data are we, the humans, along with cyber-physical and embedded systems which are monitoring both the nature (climate, environment, including humans themselves, etc.) and artificial world created by us which includes industrial processes, means of transportation and communication, software, our domiciles, factories, offices and practically everything else.

With incredible progress in embedded systems ranging from smart meters of all kinds to smartphones, we observe an explosive growth of generated data (so called "Big Data") and the fundamental challenge is how to make the Big Data small and get meaningful answers to posed questions or simplify checking validity of hypotheses. In other words, the question is how to distill out of vast amounts of raw, dumb data, the information, knowledge or wisdom. One of the keys to meaningful observations is to determine first what variables (also called features, parameters or events in different research communities) of a system or a phenomenon are most relevant or most indicative but in order to find it out, a number of preparatory steps has to be done. To reach this stage methodically, all available data should be collected first and then the process of variable (also feature) selection has to be performed in order to identify the most indicative variables as well as invariants and correlations for a given purpose.

This article poses fundamental questions regarding monitoring which we define as an automated observation and collection of data (measurements) by means of humans or sensors, meters and other devices. We need to keep in mind that a permanent strive for more data (data for the sake of data) if improperly performed may result in losing the information content that we are looking for, or even worse, lead to misinformation and wrong decisions.

Despite a large number of papers on monitoring, it still remains more of an art than a science. This article attempts to list major challenges and shed some light on how they might be tackled.

Furthermore, the collection of data is the main asset of most corporations, governments, institutions and individuals. Its quality, scope and size will have ever bigger impact on the way we think, learn, live, work, produce and create. The issue at hand is, in fact, bigger than the Big Data, as it may influence the decision making process in all walks of life including politics, economics, technology and society.

2. Monitoring Objectives

People, animals and machines observe/monitor the world with a variety of objectives that relate to the past, present or future. Inspired by Gartner's vision [1], we divide objectives into four categories with respect to time where the data are collected in order to analyze the past (what happened?), diagnose the present (why did it happen?), predict the future (what will happen?) or to construct the future (how can we make it happen?). Obviously, the ultimate goals are to understand the past and/or observe and/or control the present or control the future. The level of difficulty increases as we move along the time axis and the potential value goes up as we move from analysis and diagnosis to prediction and constructing the future. Both analysis and diagnosis can be considered as reactive methods/algorithms that are applied upon an occurrence of an event while methods and algorithms that predict and construct the future belong to a category of proactive approaches. These distinctions are summarized in Table 1.

Since observation/monitoring is the first step to almost any activity, not surprisingly it plays a pivotal role in decision making, automated manufacturing, automated driving (transportation), cooperation, consensus, adaptation and many others.

Table 1. Monitoring Objectives with respect to Past, Present and Future

Past	**Present**	**Future**	
What happened?	Why did it happen?	What will happen?	How can we make it happen?
Analyze	Diagnose	Predict	Construct
Reactive		Proactive	

For example, a consensus, that is reaching an agreement based on mutual observation and voting, can be used in many domains ranging from scheduling, decision making, resource management and reliable broadcast to reconfiguration, synchronization, fault diagnosis and fault masking.

Consequently, it is evident that monitoring plays the key role in most activities in nature and the artificial, including virtual, world created by humans. Since monitoring is so fundamental, in the next section we identify the main challenges in monitoring optimization.

3. Main Challenges

We now list five main challenges, describe current approaches and try to provide guidelines for the future.

Challenge 1: **What to monitor?**

This fundamental question depends on our application and its goal function. In today's computer we can monitor tens of thousands of its variables, probably millions in a human, but in reality we monitor much fewer of them as we are usually interested in very specific properties such as performance, reliability, security or timeliness or in case of a human a specific health condition without overloading the organism.

We are also limited by technology and the knowledge of the processes or products. We still do not have good monitors for wine but milk that is just about to turn sour, can be identified indirectly by measuring its temperature and understanding its bacteria growth process. If we have more exotic goals such as failure prediction, emotional state or privacy protection we may need actually to carry out some variable selection algorithms in order to find out what variables are the most significant ones.

A number of approaches to variable selection are well summarized in [2] while a comprehensive survey on feature selection algorithms for classification and clustering can be found in the paper by Liu and Yu [3]. In our failure prediction methodology [4, 5] we found out that feature selection has bigger impact on precision (the ratio of the number of true-positive alarms to the total number of alarms) and recall (the ratio of the number of true-positive alarms to the total number of failures) than the choice of model.

Challenge 2: **Where to place the monitors in a system? How many monitors do we need?**

The placement problem has always been a challenge: from placement (layout) of transistors on a chip to placement of nodes in wireless network for maximum connectivity. The same holds for placement of monitors.

Several solutions exist and are usually problem-specific because optimization goals vary: minimization of chip area for placement of transistors and maximum coverage for a given area with minimum number of nodes like sensors or wireless communication chips. For monitor placement, the objective is to get the most relevant data at required frequency (sampling rate) at minimum cost. Placement costs may vary significantly if, for example, some monitors have to be placed in space, underwater, underground or on a steep mountain. The complexity of a problem increases when we go from a two-dimensional to three-dimensional placement problem. The quality of placement such as coverage has direct impact on the minimum number of required sensors.

But the minimum number of monitors does not necessarily means optimum as the placement may have additional optimization criteria. How will the accuracy of the measurement be affected with larger number of monitors? Of course, cost plays a critical role and in most industrial systems the number of monitors is kept to the minimum unless an additional requirement such as fault tolerance necessitates redundancy.

There is a number of sensor placement problems and solutions [6, 7] and some of them might be adopted for the monitor placement. Specifically, for monitoring of a mobile object or a moving human, the triangulation placement method can be used which requires that each monitored object is covered by at least three sensors/monitors. The algorithms usually optimize the number of required nodes. Since the node minimization and placement problem is NP-hard, we frequently use heuristics that are not only fast but usually provide a good solution.

Additional questions that need to be asked when deciding on the placement and the number of monitors are: how reliable, secure they are, how much power they use and whether they are required to operate in real time.

Reliability of monitors should be assessed a priori as the failure of a monitor may have severe consequences. Is the design of monitoring infrastructure fault tolerant? Is it able to cope with a failure of one, two or even k monitors?

Security is another key issue that should be addressed a priori because monitor manipulation may have dire consequences. Making sure that monitoring reflects the reality under operating conditions of a system or a device is another aforementioned challenge.

Since monitors are add ons to a system operation, their power requirements must be assessed a priori and, typically, they should use as little power as possible (consider low power design) and be noninvasive.

Finally the question of time and real time is fundamental. Typically, multiple monitors need a common time base and a simple GPS-based synchronization might not be sufficient. Therefore, in some cases we need to resort to the Universal Time Coordinated (UTC) using sophisticated synchronization protocols.

Additionally, if monitors must deliver measurements in real time, meaning within given deadlines or durations, real-time system designers must ensure meeting deadlines and durations through appropriate scheduling policies and the Worst Case Execution Time (WCET) analysis.

Challenge 3: **When or how frequently to monitor?**

The question how frequently to measure a certain variable belongs to a fundamental ones and ranges from billions of samples per minute to one per day, month or a year. Fundamentally, there are three monitoring policies:

1) time triggered;
2) event triggered,
3) a hybrid.

This is a tradeoff, usually specific to a given application considering the goal, between quality of result, effect on performance and storage capability. Time-triggered monitoring requires a good understanding of a process in order to optimize the sampling frequency. Event-triggered monitoring focuses on observing changes in a system and therefore is efficient, especially in stable systems. Typically, in complex systems a hybrid approach is used as it allows to tailor monitoring of each variable according to the needs.

For example, time-triggered monitoring of an electric grid at the frequency higher than 50 Hz might not make sense because electric grid operates at 50 Hz, and therefore, the sampling period does not have to be shorter than 20 ms. The question is whether from an application perspective such monitoring improves grid's stability or not. If it takes us 40 ms to process the data then we might be forced to be satisfied with 25 Hz sampling rate as we might not be able to process the data. On the other hand, if we have two processors, we may interleave them such that we can handle a 50Hz sampling rate even with 40 ms processing time. Ultimately, the sampling frequency depends on the purpose, the hypothesis that is posed or an application.

Formally, we define the sampling rate, sample rate, or sampling frequency as the number of samples per second (or per other unit of time or event) taken from a continuous or discrete signal to make a discrete signal of a given frequency. For time-domain signals, the unit for sampling rate is hertz (inverse seconds, 1/s, s^{-1}). For example, a sampling rate for a phone is 8 kHz while High-Definition DVD requires 192 kHz. The inverse of the sampling frequency is the sampling period or sampling interval, which is the time between samples [8, 9].

Ultimately, an adaptive monitoring, especially in monitoring for prediction may help. We may change frequency of sampling of certain variables if, for example, a failure is looming in case of failure prediction or cyber attack detection.

Challenge 4: **What, where, when and how to communicate, store and process the monitoring data?**

Depending on the goal, efficiency and strategy (centralized, distributed or hybrid monitoring) the communication may require a significant bandwidth to transmit the monitoring data. This issue with questions like what, where, when and how to communicate has to be addressed in monitoring system design as requirements can be evaluated a priori.

In centralized monitoring, all monitors send the data to a single computer that is in the position to observe the status of each monitored device but also to identify trends for the entire groups of devices or monitors.

In distributed monitoring each device is monitored autonomously and it is used in the cases when communication is impossible or expensive.

The third option is a hybrid where some variables are observed locally, some might be even partially processed and then the rest is sent continuously or periodically or periodically-in-batches to a central computer. Which mode of operation to choose strictly depends on an application and user requirements.

Monitoring may produce an immense amount of data. What about how, where and when to store such data? If the data comes from multiple sources (e.g. sensors), then it is typically unstructured and arrives at different intervals. The question how this bulk of data can be stored should address the format, database organization, synchronization and storage devices. It is important to create comprehensive and expressive representation of collected data in a form of, for example, log-files that enable a flexible and semantically augmented representation of the logged events which furthermore can be analyzed automatically [10]. The next question is whether the data should be stored locally, next to monitored system or monitoring device or centrally to enable comparative analysis or a more general system view? The classical database systems are geared towards static accumulation of vast amount of data whose storage is mainly controlled by humans while monitoring systems generate data with varying frequency and/or sporadically in case of event monitoring. Furthermore, usually the latest version of data is easily accessible while earlier logs are archived. This is usually not acceptable in, for example, machine learning applications where typically the entire sequences of monitoring data are required. An example of a database, addressing most of these problems, is Aurora database [11].

Finally, we should decide where, when and how to process monitoring data. Deciding whether to process the data locally or centrally will have a direct impact on processing and communication time. It may turn out that a hybrid approach is most efficient where data are processed locally and only the relevant outcomes are passed on to a central host. Another hybrid could be that some data are processed locally and the remaining data centrally. This depends on, for example, the need for local and for global information or on processing overhead where some parts of application are processed locally and the rest is offloaded to a central server or a cloud [12]. Again and again, the methods and timing of processing the data have to be adapted to the posed questions, applications and goals.

Challenge 5: **How good is the quality of data that we get?**

Data quality is the reliance that users can put on the acquired data in terms of precision and accuracy in order to obtain a faithful reflection of monitored world. Once collected the monitoring data should remain unchanged (data stability). In the nutshell, an ideal high-quality data should be complete, adhere to standards, consistent, stable over time, accurate and time stamped. According to [13] data quality can be characterized by four attributes: accuracy, availability, interpretability and timeliness. The problem is that accuracy has many definitions. According to [13] the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The precision of a measurement system is related to reproducibility and repeatability. Additional characteristics of data may refer to completeness, severity of inconsistency (anomalies) and missing or unknown variables.

Data quality has been researched by many and good surveys can be found in [14] later in [15] and [16]. The layer of software which helps to measure and collect the data can be manipulated and can range from an obvious deception as in the VW affair to small inconsistencies that can produce a completely different picture of reality over time. This problem requires a serious considerations as what we have seen so far is a tip of the iceberg. So one of the main challenges is indeed data quality assurance.

4. Conclusions

With ever-growing hunger for data and unbounded potential of data analytics, we need to focus more on the front end of the process, namely data acquisition. This in turn requires more research on what we propose to call monitorology or the art of observing the world. Despite a lot of research and experience in automated data acquisition, several, fundamental issues remain open and methodologies tailoring a data acquisition system to any application still need to be refined.

Five challenges were posed and addressing them will make the process of distilling information from data, acquiring knowledge from information and sometimes inferring wisdom from knowledge more accurate, precise, more complete and useful. This will have in turn a significant impact on improving quality of decision making, acquiring deeper knowledge and ultimately building a better world around us.

References

[1] Gartner I T 2015 *Glossary Predictive Analytics* `http://www.gartner.com/it-glossary/predictive-analytics`

[2] Hastie T, Tibshirani R and Friedman J 2001 *The Elements of Statistical Learning*, Springer

[3] Liu H and Yu L 2005 *Toward Integrating Feature Selection Algorithms for Classification and Clustering*, IEEE Transactions on Knowledge and Data Engineering **17** (4) 491 –502
[4] Hoffmann A G and Malek M 2006 *Call Availability Prediction in a Telecommunication System: A Data Driven Empirical Approach*, Proc. 25th Symposium on Reliable Distributed Systems 83 –95
[5] Hoffmann A G, Trivedi K and Malek M 2007 *A Best Practice Guide to Resource Forecasting for Computing Systems*, IEEE Transactions on Reliability **56** (4) 615 –628
[6] Younis M and Akkaya K 2008 *Strategies and techniques for node placement in wireless sensor networks: A survey*, AdHoc Networks **6** (4) 621 –655
[7] Krause A, Singh A and Guestrin C 2008 *Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies*, Journal of Machine Learning Research **9** 235 –284
[8] Rao R 2009 *Signal and Systems*, Prentice-Hall of India Pvt. Limited
[9] Weik H M 1996 *Communications Standard Dictionary*, Springer
[10] Salfner F, Tschirpke S and Malek M April 2004 *Comprehensive Logfiles for Autonomic Systems*, Proceedings of 9th IEEE Workshop on Fault-Tolerant Parallel, Distributed and NetworkCentric Systems, Santa Fe, New Mexico, USA 211 –218
[11] Abadi D J et al *Aurora: a new model and architecture for data stream management*, VLDB **12** (2) 120 –139
[12] Fernando N, Loke W S and Rahayu W 2013, Future Generation Computer Systems **29** 84 –106
[13] Hansen D M 1991 *Zero Defect Data, PhD Diss.*, MIT
[14] Wang Y R, Storey C V and Firth P C 1995 *A framework for analysis of data quality research*, IEEE Transactions on Knowledge and Data Engineering **7** (4) 623 –640
[15] Laranjeiro N, Soydemir N S and Bernardino J 2015 *A Survey on Data Quality: Classifying Poor Data*, Proc. 21st Pacific Rim International Symposium on Dependable Computing (PRDC 2015) China 179 –188
[16] Sadiq S, Yeganeh K N and Indulska M January 2011 *20 Years of Data Quality Research: Themes, Trends and Synergies*, 22nd Australasian Database Conference (ADC 2011) Australia 235 –256

Miroslaw Malek is semi-retired Professor at the Advanced Learning and Research Institute at the Faculty of Informatics at the Università della Svizzera italiana in Lugano and consultant to governments and companies as well as expert evaluator of research projects, startup ventures and creative initiatives. Prior to that he was professor at the University of Texas at Austin (1977-1994) and at the Humboldt University in Berlin (1994-2012). He received his PhD in Computer Science from the Wroclaw University of Science and Technology in Poland. His research interests focus on dependable and secure architectures and services in distributed and embedded computing environments including failure prediction using machine learning and data analytics. His main contributions are reflected in over 250 publications and nine books. He founded, organized and co-organized numerous workshops and conferences. He served and serves on editorial boards of several journals including ACM Computing Surveys. He is Life Member of IEEE.