Learning Machines for Computational Epidemiology

Magnus Boman  
SICS Swedish ICT  
and KTH/ICT/SCS  
Electrum, SE-16429 Kista, Sweden  
Email: mab@sics.se

Daniel Gillblad  
SICS Swedish ICT  
Electrum, SE-16429 Kista, Sweden  
Email: dgi@sics.se

Abstract—Resting on our experience of computational epidemiology in practice and of industrial projects on analytics of complex networks, we point to an innovation opportunity for improving the digital services to epidemiologists for monitoring, modeling, and mitigating the effects of communicable disease. Artificial intelligence and intelligent analytics of syndromic surveillance data promise new insights to epidemiologists, but the real value can only be realized if human assessments are paired with assessments made by machines. Neither massive data itself, nor careful analytics will necessarily lead to better informed decisions. The process producing feedback to humans on decision making informed by machines can be reversed to consider feedback to machines on decision making informed by humans, enabling learning machines. We predict and argue for the fact that the sensemaking that such machines can perform in tandem with humans can be of immense value to epidemiologists in the future.

I. INTRODUCTION

Epidemiologists have enjoyed a plethora of digital tools in recent years, promising to assist with understanding the diffusion, mitigation, and real-time control aspects of communicable disease [1]–[4]. We will here broaden the perspective to distributed communicative intelligences, what we will refer to as learning machines, intelligences with a capacity to autonomously learn and communicate, with each other as well as with humans. Our goal is to identify and describe a new set of methods and technologies for assisting epidemiologists now and in the near future, going beyond state-of-the-art in artificial intelligence, taking into account recent and future developments in that research area. The use of digital tools by epidemiologists have resulted in a number of improvements to best practice [5] and through establishing a new research center for Learning Machines, we propose a new focus area for furthering cross-disciplinary work in this and related application areas.

II. METHODOLOGY

We first explain the systemic model we employ for sense-making, before attacking the celebrated problem of the micro-macro link (see, e.g., [6], [7]), here projected onto the application area of computational epidemiology. We then turn to learning machines and their importance to our suggested approach, in the form of three cases. The first two cases represent state-of-the-art while the third goes beyond, looking at what learning machines can do for the application area in the near future. Since parts of the implementation of what we propose are uncertain, we also list some of the remaining challenges for realizing the value of the underlying innovations and research results.

The long-term objective of applying intelligent data analytics to health data is to improve population health. We will focus here on intelligent data analytics, as employed chiefly by skilled human experts, serving a shorter-term objective of providing epidemiologists with improved means to understanding communicable disease. The systemic model employed rests on the proviso that the behavior of individuals is worthy of study in two ways simultaneously (Fig. 1). First, the health data internal to the body (electronic health records, medical imaging, pathology microscopy, and possibly transcriptomics and epigenomics), combined with new methods that allow for health data aggregated at the population level (and reported on in scientific publications) to be complemented by data from individuals via wide-scope health data collection and non-health data collection; so-called syndromic surveillance [8]. Data external to the human body to be analysed come in three forms (cf. [9]):

- Wellbeing (with examples including demographic change, lifestyle, behavior, and occupation)
- Environment (climate change, air quality, traffic and congestion, ambient noise, built environment, urban sprawl, waste, and sustainable food systems)
• Social (globalization of exchanges of goods and people, cultural characteristics, and socio-economic factors)

Secondly, weak and disruptive signals, including demographical trends as well as trends spotted in smaller strata of the world’s population (e.g., the increased use of wearable sensors for exercise with the purpose of rehabilitation or prevention [10]) are collected by experts in participatory processes. The latter procedure is usually referred to as networked foresight [11]. Taken together, the epidemiologist gets the status quo from the first kind of data, and predictive analytics from the second kind of data. Just bringing together data on the past, current, and future situation is not necessarily leading to better informed decisions, however. Even if context-aware computing [12] and real-time predictive analytics [13] are rapidly changing how user interaction works in many application areas, epidemiology (like many other clinical practices) rests largely on interaction with peers and on skills acquired over years of fielded practice. Therefore, ICT-support for improving upon best practice must take into account the nature of interacting experts, including the participatory aspects, of decision making in epidemiology. We believe that such practice is closer to networked foresight and scenario planning than it is to context-aware computing and advanced modeling or visualization, and that it would benefit more from the distributed and interactive intelligence features of learning machines than from incremental advances in decision analysis software (cf., e.g., [14]).

The feedback loop from human-machine sensemaking to learning machines is of the utmost importance. The meeting of minds, as well as the reasoning and decision processes that constitute sensemaking, is recurring and continuously documented. Hence, not just analyses and diagnoses are fed back into the facts database, but also meta-level and methodological information. Recent social and political developments have prompted policy makers and opinion leaders to accept the participation of individuals in healthcare decisions about them (“no decision about me without me”, see, e.g., [15]). We challenge the popular view that intelligent data analytics can in itself lead to informed decisions on complex social health issues involving individuals. Instead, we believe that a carefully controlled sensemaking process can achieve that goal.

III. THE MICRO-MACRO LINK

Complex interactions between multiple determinants of health and wellbeing (age, gender, ethnicity, lifestyle, behaviour, social and physical environment, and many other parameters) are not well understood. We meet this challenge by employing learning machines to come to terms with the micro-macro link: how population data and trends pertain to the individual [6].

Epidemiologists work with communities of practice resting on health data, and experience of the practitioners’ communities. Their practices around data usage, and the health care context in which health data resides is a very different entity from the actual data discourse. The latter consists of static stores of medical images, electronic health records, administrative and insurance records, etc. The former is a dynamic, unstable, difficult to model, and sometimes very stressful environment full of people in different occupational roles. It would be unreasonable to think that massive and exponentially growing data discourse would in its own right provide all the background information necessary for informed decisions in such a complex situation. Any model covering the data discourse must also cover its extensions into the communities of practice, and the interplay between the two. Even the smartest analytics possible is not enough: the interplay between data and practice requires finding out, for a particular stakeholder, how to best inform this decision maker based on data and the conclusions drawn from it, in a particular context. To an epidemiologist, this could mean specifying a search for correlations or causal chains of events in a data set, and waiting for an analyst to come back with results. In the near future, it could also mean entering a conference room, either in person or via mediated presence, and in that room discuss incremental real-time streaming and analysis of massive data with other people as well as with learning machines. The latter scenario possibly also enables real-time support during outbreaks or catastrophic events, an often mentioned requirement in disease surveillance (see, e.g., [16]).

To the analyst, the engineering challenge when it comes to data lies chiefly outside health data. Since most of the big data is on individuals (micro data), the promise of using micro data for analyses has a number of appealing properties when compared to using population (macro) data. Disease surveillance is performed for both communicable and non-communicable disease. Case reports (and lab reports) are essential parts of traditional disease surveillance. Syndromic surveillance adds data originally collected for other purposes, using methods that rely on detection of individual and population health indicators discernible already before confirmed diagnoses are made. Recent advances in data science have created new opportunities for syndromic surveillance, and many systems have been developed to take advantage of the new sources of data. Efficiently interpreting the combined output collected by these systems, however, remains an open problem [17]–[19]. In many cases, the populations surveyed by the systems differ significantly, complicating the application of traditional statistical (macro) methods to analyse the collected data on individuals (micro).

Arguably the most important lesson learned when addressing the micro-macro link is that many micro observations do not make up one macro observation. Even if you monitor every individual in a population and do not ever revert to sampling, what you get is aggregations of micro-observations, rather than statistics. This is as it should be, and such aggregations can be meaningful, but we wish to point out that this is not how epidemiologists typically address the problem of understanding and mitigating spread of communicable disease. Instead, they are deeply rooted in empirical and case-based work, seeking to understand current and future scenarios through the lens of earlier observations [20], [21]. This includes modeling people with homogeneous mixing and averaging under strict assumptions about uniformity (see, e.g., [22]). Big data analytics, by contrast, allows us to stick to micro observations and their aggregations, bringing in statistics only to complete observations, or to remove noise or gaps [23].
IV. LEARNING MACHINES

Today, working at scale with Big Data is almost a given when developing new analytics solutions. Available and developing technologies allow for embracing a fully data-driven perspective to an organization, where all available data is collected, analyzed, and used for informed decision making. There is no doubt an enormous commercial and societal potential in exploiting these technologies, but even though capacity to store, distribute, and analyze large data sets is there today, we argue that this is insufficient for realizing the full potential of personalized care and prevention within the health sector. In our experience from developing practical, large-scale, and autonomic analytics solutions for networked systems in general and the employment of massive register data sets for scenario planning [24], [25] and syndromic surveillance [26], [27] in particular, the current paradigm of logically centralized data collection, storage, and analytics is similarly not enough. Due to reasons of privacy, scalability, robustness, and need for autonomic configuration, we have often arrived at solutions that consist of a larger number of distributed intelligences (agents) communicating between themselves, each with a partial view of all available data, rather than one centralized solution with a complete view of the system [28].

The need for such solutions can be exemplified with the efficient implementation of mobility management and load balancing within Radio Access Networks [29]–[31]. While centralized collection of up-to-date information on the mobility and traffic patterns of all users would allow us to learn and predict behaviour, followed by the optimization of network parameters, such a solution is unrealistic due to scalability (the amount of data we would need to carry over the network would overwhelm it) and privacy concerns (the complete collection of all usage and mobility patterns in one place). Instead, we rely on simpler intelligences distributed throughout the network, learning local models of traffic and mobility by communicating between themselves and providing autonomy through local decisions.

A. Case: Watson-Like Diagnostic Systems

While current solutions exhibit simple self-organizing behavior between localized machine learning solutions, we see a similar need but on a much higher level of complexity for health applications. While one “Watson-like” [32] diagnostic system may provide high-quality diagnoses in most situations, such as second opinions on human assessments, it might be highly beneficial for this system to reason with and learn from similar systems with different knowledge bases—knowledge bases that cannot realistically be shared due to, e.g., legal and size restrictions—when arriving at a final recommendation. Further complicating the issue, while such machine-to-machine communication between intelligences may contribute to better results and competence building, humans need to be brought into a supervising role, contributing experience and early stage feedback. In other words, the distributed communicating intelligences will have to be both machines and humans. In the case of Watson, the recent launch of the cloud-based Watson Analytics platform could be a step in the right direction, assuming that the platform would allow for the more generalist intelligence required for modeling and solving problems in computational epidemiology.

B. Case: Syndromic Surveillance

Syndromic surveillance promises to deliver early signals of important events relating to communicable disease [5]. We have ourselves been involved with developing, testing, and evaluating tools for detecting early signals, for nosocomial infections [33], influenza like illnesses [34], and wider health surveillance [26]. We have also studied how policy makers collaborate with computational epidemiologists [35] and speculated on the use of collaborative intelligence for this area of application [23]. One important lesson from the community of practice is that gossip protocols for distributed intelligences might in some weak sense mirror the informal interaction between epidemiologists. A more long-term goal of syndromic surveillance can be to establish pattern-recognition algorithms for sensemaking based on multiple (and today silo-like) reporting from stand-alone surveillance systems. If this can be done in a privacy- and integrity-sensitive manner, the more generalist competence of a policy-making and/or coordinating unit for disease prevention may, again in a weak sense, be augmented by the meta-level organizing skills of a learning machine.

C. Case: Sensemaking using Learning Machines

Notions such as “the end of theory”, claiming that data can speak for itself without much need for an underlying model and that all weak signals can be found if only all data is considered, has followed in the wake of Big Data. However, an explicit or implicit model of the generating processes that can both account for noise and bias and represent higher order concepts is absolutely essential to the sensemaking process. With large numbers of diverse and high volume data streams available, it is unrealistic that such models would be fully developed and updated manually. Instead, we foresee that the representation of the data in terms of relevant features are created primarily through large scale and mainly unsupervised, learning: features that can then be used as a basis for specific supervised learning tasks such as prediction and classification. Recent developments within deep learning and transfer learning indicate that such approaches are becoming applicable to an increasing number of domains and types of data.

Such large scale learning systems will stretch the abilities of current Big Data platforms beyond their limit in terms of functionality for analysing streaming and distributed data. Current platforms are also insufficient from a machine learning perspective, which requires more flexible computational structures than provided with today’s extended map-reduce paradigm, including iteration and suitable support for the type of stateful computing that is required by continuously storing and updating an internal model of the data.

While solving the above issues would provide new levels of autonomy and information refinement, this will not in itself lead to the significant changes in practices needed to make data-driven decision making reach its full potential. Decision makers will still struggle with issues of model limitations and interpretation, as well as questions around data veracity and sample bias (cf. [36]). Again, we believe that a carefully controlled sensemaking process will be the answer. Due to restrictions exemplified in Case A and B above, this process is likely to include not only single machine intelligences,
but rather multiple distributed learning machines as well as humans in continuous interaction. Thus, one of the main challenges here is to design mechanisms for and study the behavior of such interactions.

To study these issues, we are creating the Learning Machines research center at the Swedish Institute of Computer Science. The research center and development will be focusing on machine learning at scale, both in terms of data- and representation size, and the computational platforms that support it. More specifically, we intend to study the full spectrum of communicating intelligences from simple self-organization to human interaction, and how these can contribute both to sensemaking and the creation of autonomic systems. Health applications are many and important, and computational epidemiology, which already has some track record of intense data usage and analysis, is among these. Financed mainly by industry, we intend to provide real, usable results through development of a common prototype applied on real use cases and real data acquired from several industrial and public sectors.

The generalist ambitions of Case A and Case B above will hopefully here come together in the shape of learning machines that can tackle low probability high consequence events and synthesize surveillance reports into coherent intelligence for swift action, as required by current and future health threats from communicable disease.

V. Conclusion

We conclude that an increasing amount of input from data science and from participatory networked foresight will be made available to epidemiologists in the future. Through computational epidemiologists and learning machines being situated at the clinics and at communicable disease surveillance units, a process of sensemaking can take place. The output of that process allows for epidemiologists to take better informed decisions and so to complement or even improve upon current best practice. This contributes to the long-term objective of improving health for all. Challenges include:

- Explaining the value of the output of sensemaking, thus motivating the time and effort invested by epidemiologists.
- Creating opportunities for sensemaking, via mediated presence and dialog (human-human, human-machine, and machine-machine).
- Customizing learning machines for the communicable disease domain, as well as creating generalist and meta-level intelligence.
- Ensuring communication, learning, and decision stability in large networks of adaptive and (at least partially) autonomous entities. Complex networks of communicating intelligences could potentially suffer from oscillations or catastrophic events in terms of the number of communicated queries between them, continuous bias confirmation in learning and feedback, and both globally and locally sub-optimal decisions. As the systems we envision could be orders of magnitude more advanced in terms of scale, adaptivity, and autonomy as compared to existing control systems, for example, traditional methods of evaluating system stability are likely insufficient. Devising new methods [37] for studying and ensuring stability and performance of these complex systems is therefore essential.

Our micro-macro link analysis in computational epidemiology can be summarized as a reminder that many micro observations do not make up one macro observation. This holds true also for the complex networks that we have analyzed for other application areas than health [38]. When analyzing complex networks with self-organizing capabilities, as we have done for several industrial clients, one may view the entire network as a macro representation of a dynamic system state. A micro representation, by contrast, would only provide a myopic view of a local state. Any visualization of the entire system is traditionally seen as motivating a systemic view, but in our experience, series of snapshots of typical patterns can be very informative. If the network is equipped with distributed intelligence, it is this myopic view that constitutes the boundaries of what decisions may be taken. Thus, true autonomy of a complex network makes more sense when decentralized rather than centralized. We believe the same argument holds for learning machines.

Acknowledgment

The authors would like to thank Ericsson AB for their continuous support for developing distributed intelligence and autonomous networks, and EIT ICT Labs for their push for methodological development and use of networked foresight.

References


