Public Health and Epidemiology Informatics

A. Flahault[1,2], A. Bar-Hen[2], N.Paragios[3,4]
[1] Institute of Global Health, University of Geneva, Geneva, Switzerland
[2] Centre Virchow-Villermé, Université Paris Descartes, Paris, France
[3] Center for Visual Computing, CentraleSupélec, University of Paris-Saclay, Saint-Aubin, France
[4] Inria, France

Summary

Objectives: The aim of this manuscript is to provide a brief overview of the scientific challenges that should be addressed in order to unlock the full potential of using data from a general point of view, as well as to present some ideas that could help answer specific needs for data understanding in the field of health sciences and epidemiology.

Methods: A survey of uses and challenges of big data analyses for medicine and public health was conducted. The first part of the paper focuses on big data techniques, algorithms, and statistical approaches to identify patterns in data. The second part describes some cutting-edge applications of analyses and predictive modeling in public health.

Results: In recent years, we witnessed a revolution regarding the nature, collection, and availability of data in general. This was especially striking in the health sector and particularly in the field of epidemiology. Data derives from a large variety of sources, e.g. clinical settings, billing claims, care scheduling, drug usage, web based search queries, and Tweets.

Conclusion: The exploitation of the information (data mining, artificial intelligence) relevant to these data has become one of the most promising as well challenging tasks from societal and scientific viewpoints in order to leverage the information available and making public health more efficient.

Keywords

Big data, learning machine, data analytics, pharmaco-epidemiology; disease surveillance

Yearb Med Inform 2016:240-6
http://dx.doi.org/10.15265/IY-2016-021
Published online November 10, 2016

Introduction

"*Every 50 years, there is a revolution in healthcare based on the trends of the era. In the 1870s, healthcare was revolutionized by the germ theory of disease and promotion of public health efforts. In the 1920s, the discovery of penicillin propelled forward the use of medication as treatment for disease. In the 1970s, use of the randomized controlled trial (RCT) ushered in an era of evidence-based medicine. As we approach the 2020's, the trend toward big data, tools, and systemization of care will revolutionize the way hospitals and physicians work and, most importantly, the way patients are treated…*" [1]. This statement confirms that public health is transitioning to more data-driven policies for efficiency, cost-savings, equity, and improved outcomes. To achieve these goals, big data is acclaimed to be the most efficient evolution vehicle in this decade. There are indeed a variety of factors that have converged in the past few years, which increase the amount of digitized healthcare information:

- The convenience of electronic health records has increased the number of hospitals and providers who use them, subsequently increasing the amount of electronic data generated.
- Social security and medical insurances need large amounts of information to be analyzed in order to more accurately understand what occurs to patients.
- New technology in general, including devices, implants, and mobile applications on smartphones and tablets, has focused on improving healthcare services and contributed to the increased amount of data available to providers.

There is a high pressure on clinical practices to become evidence-based and predictive. Providers, patients, and the entire healthcare industry [2] realize a variety of beneficial implications from the use and analysis of big data. Without an initiative to aggregate, manage, and analyze big data, the public health arena would be suffering from information overload. Big data research has emerged in the recent years due to the proliferation of data management systems along with tremendous progress made over the past decade in computing power. Furthermore, on top of computing progress, scientific progress has been made in parallel that has contributed to the creation of a new discipline at the intersection of computer science and applied mathematics (namely statistics, machine learning, optimization), the data science domain. The main domain objective is to develop mathematical models and computational solutions able to reason and interpret massive amount of data where typically the information sought is quite sparse. The complexity of the task is mostly due to three important challenges: (i) the dimensionality of data/observations that is often huge, with data/observations of various nature and heterogeneity, (ii) the sparse nature of critical events where one should be able to determine/develop solutions to events that appear in a non-uniform and rather non-frequent manner, and (iii) the volume of the measurements in terms of samples/observations where a colossal amount of data is collected in a continuous fashion.

Our paper is organized as follows: first, we provide a short review of the state of the art of data mining methods; second, we propose some possible theoretical directions specific to the field of health science; and

third, we present recent applications and examples of how data science is transforming the field of epidemiology and public health.

Challenges and Perspectives in Data Science

Data mining on big data is achieved through either unsupervised or supervised methods. In the former case, the objective is to identify common behaviors/characteristics within the observed population in the absence of "ground truth" knowledge. Supervised methods rely on a different principle seeking to reproduce observed behaviors. The idea is, given a set of observations and their "corresponding" behavior/final outcome, to determine an algorithm that can actually reproduce the outcome. Large sets of heterogeneous data are difficult to interpret. This is mostly due to two reasons: (i) first, unsupervised mining methods become quite unstable (e.g. defining the appropriate metric to compare examples becomes infeasible), and may fail to provide any meaningful interpretation, while (ii) supervised/learning methods may also fail because the expected ratio between training examples and dimension of the data is not satisfied.

Efficient Representations of Big Data

In typical data science examples, the first step consists of reducing the dimension of data while preserving the ability to reason on them. This can be done using well known and quite standard statistical hypotheses/models or non-linear/data-driven reduction methods. Linear approaches like for example Principal Component Analysis (PCA), Independent Component Analysis (ICA), Non-Negative Matrix Factorization [3] are some of the most commonly used tools to reduce dimensionality of data. The central idea is to determine a linear base (with base vectors of the same dimension as the one of the original observations) with a small number of elements and then represent the observations through the coefficients of the projections to this base. Canonical Correlation Analysis [4] seeks to determine the transformation that would create the best possible correlations between the individual variables of the observation factor. The dimensionality of the problem then drops to the number of the base elements. These methods work well when observations satisfy base assumptions like for example following a normal distribution, exhibiting linearity, and in the most general case, referring to observations of similar nature. In the recent years, we witnessed a new research direction that refers to a data-driven learned non-orthogonal base due to the compressed sensing approach. It determines a base where the projection of observations generates a sparse matrix [5]. This was further explored to encode the concept of group/structured sparsity [6] where the projection of observations in groups of variables is sparse rather than the individual ones.

Handling heterogeneous and non-linear data is a more challenging task. The most trivial approach consists in introducing a projection on the observation space that will take the non-linear data and project them to a linear space using for example kernels. This can be seen as a transformation through an inner product between the kernel and the observations that will create a linear projection. Once this has been achieved, conventional dimensionality reduction methods can be used to reduce the complexity of the observations [7]. This principle works well assuming that the right kernel can be identified which is highly complex in most cases – in particular when the dimensionality of the original data is high. Geometric methods/embeddings are a direct alternative to deal with non-uniform data. They consist of approaching data as a manifold endowed with a distance and then reasoning on this new manifold through geometric/graph-based reduction methods. Multi-Dimensional Scaling [8] aims at projecting data to a new space where distance between two observations is preserved. Isomap [9] endows multidimensional scaling with local approximation of manifold in order to determine distance as a minimum geodesic path relating the two observations. Locally Linear Embedding uses similar principles as discussed before with the exception that samples/observations are now determined as a linear combination of their neighbors. Laplacian Eigen maps [10] better encode intrinsic geometry of data through the construction of neighborhood proximity graphs where Eigen functions of Laplace–Beltrami operator are applied on the manifold to produce embedding space and new dimensions of data.

The outcome of such reduction methods is the projection of a high-dimensional, non-interpretable observation set to a low dimensional manifold. Such a process introduces two important benefits in mining observations. First, it will allow human interpretation. Low dimensional manifolds can indeed be more easily visualized and allow humans to find correlations between expert knowledge and mathematical intelligence. Second, it will allow the development of efficient and robust predictive algorithms and behaviors that can be further generalized to unobserved data. This is due to the fact that the number of degrees of freedom of data has been decreased and therefore, prediction methods are less sensitive to over-fitting. Once data have been projected to a lower dimensional space, then the next step consists of reasoning on it, a process that is done either in an unsupervised or a supervised manner.

Machine Learning Paradigm

Unsupervised mining consists in developing methods that are able to separate low dimensional samples to as many as possible distinctive classes. Supervised classification associates samples with latent variables (expected outcomes of the prediction mechanism) and then seeks methods that are able to separate new samples with a performance comparable for what was observed for training.

Unsupervised Classification Methods

Unsupervised classification methods mostly rely on the statistical interpretation of data. Observations are considered to form a distribution at some arbitrary space and what are sought are the modes and parameters of these distributions. Primitive approaches to address this task refer to clustering methods. K-means and its numerous variants are the simplest solution to this problem. Given

the number of hypotheses and the initial solutions, this method recursively updates partitions, such that samples are optimally distributed according to cluster-independent Gaussian hypotheses. More complex clustering methods seek to eliminate the need to determine the number of a priori clusters and to this end, approaches like mean shift mode [11] seeking and clustering, affinity propagation [12], or more recently graph-based methods based on linear programming were introduced [13]. Bayesian classification is a step forward where samples are considered to form a multivariate density function. Mixture models (with different support kernels like Gaussians/Laplacians, etc.) are then employed to determine behaviors with respect to different hypotheses. Parameters of these models are determined using the expectation-maximization principle [14]. The main strength of these methods is its simplicity while their main limitation is the implicit "uniform" assumption on the distribution of samples with respect to different classes, this assumption being often violated when seeking to determine abnormal behavior / rare events in the context of data sciences.

Supervised Classification Methods

Supervised methods are more efficient in developing prediction mechanisms able to interpret massive data. The central idea is to build a database where measurements are associated with "expected" outcomes and then seek to construct a mathematical model that is able to reproduce the expected outcomes given the measurements for this set. Once this mechanism is build, then it can be applied as it is to new measurements towards predicting their class/origin. Early approaches in this direction refer to logistic regression and support vector machines. Boosting [15], the term that is often employed for logistic regression, seeks to determine a set of "linear" weak classifiers that once combined are expected to produce a stronger classifier. Support vector machines [16] are similar to boosting but the aim is different: support vector machines provide a hyper-plane built on a high dimensional space, which allows for separating different classes. This hyper-plane is determined according to a set of support vectors that are learned from data. Random forests [17] as alternative to the aforementioned methods follow the same principle. Given a set of training examples and associated decisions, they construct binary decision trees. For example, given a new sample and using binary decision principle, they propagate the new sample to the appropriate leaf. In order to remove the over-fitting bias, multiple trees can be constructed and then combined for optimal decisions. Such methods inherit simplicity, computational efficiency, and provide a good tradeoff between performance and difficulty to use/employ/learn. On the other hand, decisions are often taken individually and no correlation between variables/measurements is considered.

Structure prediction [18] is a step forward in introducing context in decision process, either through correlation between variables or through correlation between examples. Instead of building individual classifiers, what is sought here is a classifier able to reason at a larger scale in combining samples and observations. More structured representations like Markov Chain [19] and Hidden Graphical Models are also used for the same purpose. By training the graphical model of relatively simple structures, the same task of classifying new observations given what is measured is performed. Probabilistic graphical models [20] are a class of methods aiming at reasoning over graphs. The classification problem is often expressed as an optimal labeling one over the graph nodes. This graph inherits the connectivity that is often used to impose consistency and measure correlation between different variables and principles like max flow-min cut are used to determine the optimal assignment for a new sample. Artificial neural networks [21] are a particular case of graphical model inheriting some biological interpretation where, given a decision structure, what is learned is the way how decisions are propagated over these networks to optimally express behaviors of training data. These methods gained interest in recent years due to the development of efficient learning algorithms, their computational solutions along with the availability of computing power for parameters optimization at training (namely deep learning [22]).

Next we will discuss some concrete examples on the interest of using such methods in the context of health and epidemiology.

Big Data: Applications to Public Health and Epidemiology

Data Science for Disease Surveillance

Previously, when medical records were mainly paper-based, it could take weeks to find out that an infection was emerging somewhere in the world. Nowadays, most US hospitals use electronic medical records, and the growing prevalence of electronic medical records has had an unexpected benefit: by combing through data now received almost continuously from hospitals and other medical facilities, some health departments are spotting and combating outbreaks at an unprecedented speed. In February 2012, public health officials in Michigan noted an increase in electronic reports from clinical laboratories indicating E. coli cases in several counties. In less than a week, officials had enough evidence to alert the public about the infection probably linked to clover sprouts in food at the Jimmy John's sandwich chain. The chain quickly removed the sprouts and by April, the outbreak had died [23]. France was a precursor with the INSERM "*Sentinelles*" network [24], a fully electronic network based on a network of general practitioners throughout France. The "*Sentinelles*" network was set up in 1984 by Alain-Jacques Valleron, using at this time the Minitel [25], an early online service. New health technologies, including the innovative use of the routinely assessed data, in strict compliance with national regulations regarding confidentiality and data protection, should enable health authorities to respond readily to health crises and offer great potential for significant growth in this sector.

Ginsberg et al. [26] and Polgreen et al. [27] published two independent studies showing that the large number of queries on Internet search engines (*Google* and *Yahoo*) could be used to detect influenza epidemics in the USA one to two weeks earlier and five weeks ahead of the real time mortality ob-

served by the Atlanta-based US *Centers for Disease Control and Prevention* (US CDC). A set of 45 search queries monitored on Google (as "*indications of flu*") providing in real time a means of flu monitoring all over the USA were made publically available. The authors state that this early warning system was not in direct competition with the systems based on sentinel doctors or virology laboratories: once the alarm has been raised it was still necessary to check on the ground whether it was influenza and, if so, what was the strain, etc. Let's consider in more details the controversy recently raised after the flu season in New York City in Winter 2013. "*When influenza hit early and hard in the United States this year, it quietly claimed an unacknowledged victim: one of the cutting-edge techniques being used to monitor the outbreak*", wrote Declan Butler in Nature [28]. A comparison with traditional surveillance data showed that Google Flu Trends, which estimated prevalence from flu-related Internet searches, had drastically overestimated peak flu levels. The glitch is no more than a temporary setback for a promising strategy, experts said, and Google thought initially to refine its algorithms but finally closed their website, in August 2015. But flu-tracking techniques based on web data mining and on social media proliferated, and the episode should be considered as a reminder that these methods may complement, but not substitute traditional epidemiological surveillance networks. "*It is hard to think today that one can provide disease surveillance without existing systems*," said Alain-Jacques Valleron. "*The new systems depend too much on old existing ones to be able to live without them,*" he added. Google Flu Trends has continued to perform remarkably well, and researchers in many countries have confirmed that its influenza-like-illness (ILI) estimates were accurate. But the 2013 US flu season seems to have confounded Flu Trends algorithms. Estimates for the Christmas national peak of flu were almost twice the CDC's, and some state data showed even larger discrepancies. It was not the first time that a flu season had tripped Google up. In 2009, Flu Trends had to tweak its algorithms after its models badly underestimated ILI in the United States at the start of the H1N1 (swine flu) pandemic — a glitch attributed to changes in how people were searching the net using Google as a result of the exceptional nature of the pandemic [28]. So these algorithms should continue to be tweaked to account for changing search trends and those information seekers, who are merely responding to an increased awareness of risk as well as those, who are ill and seek medical information related to their symptoms.

We can be sure that the world's health surveillance will use this type of approach in addition to the systems currently set up, especially in the increasing number of places where Internet is accessible and used. Many experts welcomed a major advance in terms of early warning that should help provide wider coverage, more precise monitoring, and in the long term better understanding of the dynamics of epidemics and the conditions that lead to their outbreak. Weather forecasting was revolutionized first by setting up a worldwide network of ground sensors and second by the introduction of satellite communications. What was clearly lacking in health monitoring until now was the missing links in the chain of sensors (doctors or laboratories), too widely spaced, too imprecise, relying on good will and volunteers, and subject to under-reporting. The use of Internet search engines, that could guide operations to look into possible sources of epidemics in specific places around the world, opens up the way to a re-organization of health surveillance and the reappraisal of the role played by those involved.

Some new problems are emerging: will this data (derived from the search engines) that is now free and available almost in real time in many areas of the world remain freely available? Will there be platforms that will analyze the data using standardized, recognized, and reliable methods? The set of appropriate search queries for the optimum detection clearly varies according to culture, language, customs, and health systems: the impact of these variations on the quality of the estimates should be studied and the best set for each country should be defined and monitored to see how the performance changes (sensitivity and specificity of diseases detection). Will the poorest areas in the world be covered by this technological advance (perhaps through mobile devices) or will they, once again, be left by the wayside, bearing in mind that many of these areas are hot spots in terms of outbreak of infectious diseases of pandemic potential such as Ebola or Zika virus disease.

Similarly to infectious diseases, there is an urgent need for surveillance systems for chronic diseases such as cardiovascular diseases or cancer. The analysis of trends in chronic diseases is often based on mortality data which do not pick up early changes in disease morbidity. Data from registries or surveys may show earlier changes in trends. However, for many chronic diseases, registries exist only in certain areas and may not be representative of the whole population. They are also often subject to a lack of adequate resources (*e.g.*, financial resources). Existing routinely assessed data may provide the additional information required to implement an early warning system, for example as to the potential increase of a medication side effect. Hospital discharge data, data from health insurance companies or emergency medical services, data on drug sales etc. all provide valuable additional information that should have an impact on the planning of public health interventions to tackle threatening epidemics. Successes in public health and medicine such as the treatment of H. pylori infection to prevent gastric cancer demonstrate that trends in infectious diseases and chronic diseases should be analyzed in a combined way. Especially, big data provides the unique opportunity to discover early links between diseases to advance research further. Explorative findings in the analyses of big data will instigate basic and epidemiological research for confirmation. The translation into public health following the evaluation in comparative intervention studies as well as the subsequent setting- and population-based evaluation will be a major tool for improving health outcomes. The success of public health interventions will then be monitored again using big data to form a truly interdisciplinary and integrative cycle of research and implementation. This approach will be innovative, timely, and allows for improving the health of the population on a large scale.

Big Data for Drug Safety and Computer-Assisted Pharmacovigilance

It takes often many years after a medication has been put on the market before the medication is found to have caused an undesirable event. This was the case for distilbene (vaginal cancer), phenacetin (chronic renal failure), amidopyrin (medullary aplasia), isomeride and, more recently, mediator (valvulopathy), to give but a few examples from a far longer list of drugs that are sometimes less well known. Pharmacovigilance is the careful analysis of reports by doctors to public health authorities and/or pharmaceutical companies. Pharmacovigilance measures were set up in the USA and in Europe after the regrettable case of thalidomide which caused serious foetal malformation (phocomelia) in the early 1960s. Nevertheless, the delays between the time a new medicament is put on the market, the time of the discovery and then the confirmation of an adverse drug event (ADE) caused by the medicament, and the design of appropriate regulations to prevent such events are still far too long. These delays, which result in morbidity and mortality that could be avoided, are considered by patients and consumers as a sign of failure of the public health system and contribute to their loss of confidence in health authorities.

In a study using data drawn from queries on Google, Microsoft, and Yahoo search engines, authors have for the first time [29] been able to detect evidence of unreported prescription drug side effects before they were found by the Food and Drug Administration's Adverse Event Reporting System [30]. This study is based on data-mining techniques. Authors used automated software tools to examine queries handled by six million Internet users taken from Web search logs in 2010. They looked for searches relating to an antidepressant, paroxetine, and a cholesterol lowering drug, pravastatin. They were able to find evidence that the combination of the two drugs caused high blood sugar.

The scope of existing pharmacovigilance systems is limited by the fact that evidence could be generated only when physicians notice something and report it. Researchers turned to computer scientists at Microsoft, who created software for scanning anonymous data collected from a software toolbar installed in Web browsers by users who permitted their search histories to be collected. Scientists were able to explore 82 million individual searches for drugs, symptoms, and conditions. They first studied individual searches for the terms paroxetine and pravastatin, as well as searches for both terms during 2010. They then computed the likelihood that users in each group would also search for hyperglycemia as well as roughly 80 of its symptoms — words or phrases like "*high blood sugar*" or "*blurry vision*" [31]. They stated that people who searched for both drugs during the 12-month period of the study were significantly more likely to search for terms related to hyperglycemia than those who searched for just one of the drugs. They also found that people, who did the searches for symptoms relating to both drugs were likely to do the searches in a short time period: 30% did the search on the same day, 40% during the same week, and 50% during the same month for both drugs.

The strength of the signal authors detected in the searches was high, and it will be a valuable tool for official drug agencies to add to their current systems for tracking adverse effects. "There is a potential public health benefit in listening to such signals," they wrote in the paper, "and integrating them with other sources of information." In the future, researchers could add new sources of information, like behavioral data and information from social media sources. One challenge will be to integrate new sources of data while protecting individual privacy.

Cami A. et al., 2011 [32] described the results of a method that could radically change standard drug safety. It is a predictive approach that can determine new ADEs before they have occurred. The authors constructed a network representing drug-ADE associations for 809 drugs and 852 ADEs on the basis of a snapshot of a widely used drug safety database from 2005. The data came from various sources: pharmacovigilance reports, pharmaceutical taxonomies, ADE taxonomies, and the intrinsic pharmacological properties of the drugs studied. The authors trained a logistic regression model to predict unknown drug-ADE associations that were not listed in the 2005 snapshot. They then compared these predictions with the new drug-ADE associations that appeared in a 2010 snapshot of the same drug safety database (2006-2010). The accuracy of the predictions was amazing. Using this method, the authors were able to predict with a high specificity 7 of the 8 ADEs that emerged after 2005, including the association between the anti-diabetic rosiglitazone (Avandia) and the occurrence of heart attacks.

As the editors commented on the article, there are benefits for patients. With this powerful model in place, certain unknown ADEs could be predicted, which helps to prevent morbidity and mortality related to suspect drugs by providing more appropriate information to consumers.

Both examples detailed above seem to herald a new era with the need for new skills for drug safety teams, both in the industry and the public domain: the application of this type of methods is not trivial and requires extensive skills in the domains of computing and statistics that were not required by the standard monitoring of ADEs (which requires more knowledge of pharmacology and medicine). There is no doubt that this kind of software will now be produced as integrated systems and that it will be difficult not to use them to focus the attention of pharmacovigilance teams more closely on the risks identified as having a high probability of occurrence. However, computers will not replace human-made pharmacovigilance but they could be of considerable assistance.

Big Data for Tracking Nosocomial Infection and Reducing Hospital Readmissions

In hospitals, data science is changing the way physicians take care of patients at the individual level, fostering more personalized support right at the patient's bedside. For instance, NorthShore University Health System in Evanston, Illinois, USA, has observed the impact of predictive modeling at the point-of-care [33]. As a result of its large data sets, the health system has developed models to identify which patients are likely carriers of a dangerous microorganism, Methicillin-Resistant Staphylococcus Au-

reus (MRSA). By implementing the results of that modeling into electronic medical records, providers within the health system receive alerts when a patient is admitted that meets the characteristics of being a high-risk carrier of MSRA, as determined by the predictive models. These models are able to identify about 90% of MRSA in the patient population. NorthShore has also used this modeling to predict which patients are likely to develop Clostridium difficile.

Big data also provides predictive models for the likelihood of readmission within 30-days [34]. A user can look at a panel of patients to see which patients are at risk — high, medium or low — of being readmitted in 30 days. These predictive models which identify patients, who were recently discharged from the hospital with a high risk for readmission, are able to send messages to primary care practices. Without large data sets showing trends and patterns in huge groups of patients, this type of accurate predictive modeling would not be possible. Big data is emerging in the healthcare space, and it is likely that it will continue to magnify over time. Healthcare organizations are going to keep collecting massive volumes of data, so aggregating and analyzing that data will be a continual challenge. However, that effort will be worthwhile as we begin to see the fulfillment of big data promises.

Data & Model-driven Approaches in Health/ Epidemiology Sciences

The afore mentioned examples demonstrate the interest in considering data-driven approaches towards the better understanding of health-related events or the more precise prediction of future ones. However, health and epidemiology sciences can be considered parts of the domains where machine learning/data science methods would be applied. The advantage of such view point is that algorithmic development is separated from the domain, and adoption of algorithms becomes a question of performance. Such an appealing perspective however contradicts the underlying philosophy of health sciences that rely on expert knowledge and models, and most of all requires human understandable and interpretable solutions. The recent progress made in machine learning and optimization should be used as the driving force to bridge the gap between pure data-reasoning and model-driven approaches. It is time to associate models exploiting basics of human intelligence and interpretation with the power of data collection, annotation, and diversity.

Graphical models are an appropriate tool to introduce context and relate algorithmic suggestions with human interpretable solutions. Deep learning [35] is a method that exploits full potential of data and unlocks their hidden intelligence. Unlike the two learning paradigms discussed earlier, namely unsupervised learning and fully supervised learning methods, state of the art methods should reflect the true availability of data in real life. Specifically, real-world data is often partially supervised, that is, while the training examples are annotated, their annotations contain missing information. We propose to represent this missing information as latent variables of graphical models, which are neither observed during training nor tested.

Given a set of weakly supervised training examples and over-complete representation in respect to the variables interaction, we need to determine the best possible model in terms of compactness and prediction behavior. In order to deal with the large amount of data, we will put the emphasis on the progressive/online learning concept where the model is progressively built and updated using new examples/labels. In order to deal with the inherent noise in weakly supervised data, we still need to investigate principles from self-paced learning and active learning among others.

References

1. Riskin D. The Next Revolution in Healthcare. Available online at http://www.forbes.com/sites/singularity/2012/10/01/the-next-revolution-in-healthcare/. Last accessed 7/28/2016.
2. Herper M. Forbes Healthcare Summit: Using Big Data To Make Patients Better. Available online at http://www.forbes.com/sites/matthewherper/2013/02/05/forbes-healthcare-summit-using-big-data-to-make-patients-better/#a4e34f33e4a1. Last accessed 7/27/2016.
3. Lee D, Seung S. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788-91.
4. Knapp T. Canonical correlation analysis. A general parametric significance-testing system". Psychological Bulletin 1978;85(2):410–6.
5. Candes E, Romberg J, Tao T. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics 2006;59 (8):1207.
6. Huang J, Zhang T. The benefit of group sparsity. The Annals of Statistics 2010;38(4):1937-2586.
7. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 1998;10(5):1299-319.
8. Borg I, Groenen P. Modern Multidimensional Scaling. Theory and Application. New York, NY: Springer; 2005.
9. Tenenbaum J, de Silva V, Langford J. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 2000;290(5500):2319-23.
10. Belkin M, Niyogi P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Adv Neural Inf Process Syst 2001;14:586–691.
11. Cheng Y. Mean Shift, Mode Seeking, and Clustering. IEEE Trans Pattern Anal Mach Intell 1995;17(8):790–9.
12. Frey B, Dueck D. Clustering by passing messages between data points. Science 2007;315:972–6.
13. Komodakis N, Paragios N, Tziritas G. Clustering via LP-based Stabilities. Adv Neural Inf Process Syst 2009;21:865-72.
14. Duda R, Hart P, Stork D. Pattern Classification. Wiley-Interscience; 2000.
15. Freund Y, Schapire R. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. J Comput Syst Sci 1997;55(1):119-39.
16. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273.
17. Breiman L. Random Forests. Mach Learn 2001;45(1):5–32.
18. Tsochantaridis I, Joachims T, Hofmann T, Altun Y, Singer Y. Large margin methods for structured and interdependent output variables. J Mach Learn Res 2005;6(9).
19. Norris J. Markov Chains. Cambridge University Press; 1998.
20. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press; 2009.
21. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1994.
22. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798-828.
23. Freudenheim M. Fast Access to Records Helps Fight Epidemics. Available online at http://www.nytimes.com/2012/06/19/health/states-using-electronic-medical-records-to-track-epidemics.html?_r=0. Last acceded 7/28/2016.
24. Sentinelles, the French Network for Electronic Surveillance on Communicable Disease. Available online at http://websenti.u707.jussieu.fr/sentiweb/. Last acceded 7/28/2016.

25. Minitel. Available online at https://en.wikipedia.org/wiki/Minitel last acceded 7/28/2016. Last acceded 7/28/2016.
26. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012-4.
27. Polgreen P, Chen Y, Pennock D, Nelson F. Using internet searches for influenza surveillance. Clin Infect Dis 2008;47(11):1443-8.
28. Butler D. When Google got flu wrong, US outbreak foxes a leading web-based method for tracking seasonal flu. Nature 2013 Feb 14;494(7436):155-6.
29. White R, Tatonetti N, Shah N, Altman R, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013;20(3):404-8.
30. US Food and Drug Administration. Questions and Answers on FDA's Adverse Event Reporting System (FAERS). Available online at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. Last acceded 7/28/2016.
31. Markoff J. Unreported Side Effects of Drugs Are Found Using Internet Search Data, Study Finds. Available online at http://www.nytimes.com/2013/03/07/science/unreported-side-effects-of-drugs-found-using-internet-data-study-finds.html?_r=0. Last acceded 7/28/2016.
32. Cami A, Arnold A, Manzi S, Reis B. Predicting Adverse Drug Events Using Pharmacological Network Models. Sci Transl Med 2011;3(114):114-27.
33. Robicsek A, Beaumont J, Wright M, Thomson R Jr, Kaul K, Peterson LR. Electronic prediction rules for methicillin-resistant Staphylococcus aureus colonization. Infect Control Hosp Epidemiol 2011;32(1): 9-19.
34. Roney K. The Rise of Big Data in Hospitals: Opportunities Behind the Phenomenon. Available online at http://www.beckershospitalreview.com/healthcare-information-technology/the-rise-of-big-data-in-hospitals-opportunities-behind-the-phenomenon.html. Last acceded 7/28/2016.
35. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

Correspondence to:
Prof. Antoine Flahault
Chair Louis Jeantet in Public Health
Director of the Institute of Global Health
Faculté de Médecine, Université de Genève
Campus Biotech, Cheimin des Mines 9
1202 Geneva, Switzerland
E-mail: antoine.flahault@unige.ch