Health Actuaries and Big Data

December 2017

International Actuarial Association | Association Actuarielle Internationale

This is a paper of the Health Committee of the IAA and has been produced and approved by that committee.

International Actuarial Association
Association Actuarielle Internationale
99 Metcalfe Street, Suite 1203
Ottawa, Ontario
Canada K1P 6L7
www.actuaries.org
Tel: 1-613-236-0886   Fax: 1-613-236-1386
Email: secretariat@actuaries.org

Introduction

Data is the basic resource used by actuaries and other analysts for modelling, pricing, risk adjustment, and reserving of health insurance plans, as well as in epidemiological/pandemic studies in public health management.

Traditionally, due to computational limitations, the data used was limited in scope and structure; it required use of "traditional" (usually parametric) statistical techniques to enable analysis and prediction with the available computational resources. "Big Data", in contrast, became available only in recent decades due to the enormous increase in computing power, networking, cloud computing, and storage. Big Data is usually characterized by voluminous amounts of data from various sources, often unstructured, wherein are hidden many phenomena and trends—thus forcing the user to "mine" and identify important or interesting trends. The computational resources currently available enable use of simulations and other non-distributional techniques and support real-time processing to dynamically change performance and actions (as is often done in marketing).

Big Data is an extension of data resources and analysis tools that have traditionally been used by analysts and actuaries. This extension gave birth to a new profession (data scientist) that uses specific skills in analyzing and delivering actionable insights from data in general and Big Data in particular. Drew Conway (2013) defines data scientists as people with skills in statistics, artificial intelligence, machine-learning algorithms, data mining, and (Big) data management and processing. Machine learning automates analytical model-building by using algorithms that iteratively learn from data, thus allowing computers to find hidden insights without being explicitly programmed where to look (SAS, 2016). The rapid growth in models and computing power has allowed data scientists to move into areas that have traditionally been the responsibility of other professionals (such as actuaries), resulting in a need for actuaries to acquire and understand these new skills quickly and use them to emphasize their differentiation in the market.

What is Big Data?

Big Data does not only refer to very large datasets. It is typically understood to refer to high volumes of data, from many sources and often different time periods, requiring a high velocity of ingestion and processing and involving high degrees of variability in data structures (Gartner, 2016). Given this, Big Data involves large volumes of both structured and unstructured data (the latter referring to free text, images, sound files, and so on) from many different sources, requiring advanced hardware and software tools to link, process, and analyze such data at great speed. In healthcare, we are used to using very large datasets—bigger than those used in life, pension, or general insurance practices. Some other practice areas refer to "Big Data" sets that are relatively small by healthcare standards.

Data is rapidly increasing in volume and velocity because of developments in technology, involving many more sensors, networking, and cloud storage, which constantly generate, store, and distribute streams of data. Examples of these include fitness or wellness tracking devices, car tracking devices, patient tracking in hospitals, and medical equipment. The expansion of cheap data storage has permitted the development of applications that allow the storage of large volumes of self-reported data. People contribute to the rapid expansion of data, primarily in the form of social media and online interactions.

Self-reported data presents its own problems for analysis: it is often incomplete and un-validated, and can suffer from self-selection bias and potentially contradict data from other sources.

IBM estimates that at least 80 per cent of the world's data is unstructured (Schneider, 2016), in the form of text, images, videos, and audio. This data may contain valuable unique insights for an organization, enabling it to more effectively meet customers' needs and answer queries in real time, among other applications; being unstructured, it requires new processing and analysis techniques. This data environment is very different from the sets of structured data that actuaries and other analysts have traditionally used, and requires investments in hardware, software, processing processes, education, and skills.

The increase in volume, velocity, and variability of data has increased the demand for processing and storage power, and Big Data typically cannot be stored and analyzed in traditional systems. To handle Big Data, organizations typically have to introduce large-scale parallel processing systems. This allows them to store vast amounts of data of all types on low-cost commodity hardware, or in cloud storage, and query and analyze the data in real time by parallelizing operations that were previously done on a single or few processors.

If an organization implements systems that enable it to access and store large quantities of data, that is, however, only the first step. According to Gary King (2016), "*Big Data is not about the data*"—while the data may be plentiful, the real value emerges from the analysis of this data and, beyond that, a responsive operational environment that allows for the application of insights, often in real time. For example, the volume of sensor-based data creates major problems of distinguishing signal from noise. As actuaries, we have training (and experience) that helps us to do this with traditional data sources. We are able to distinguish, for example, when conditions warrant changing pricing or reserving assumptions. Although actuaries may analyze Big Data, they often lack the skills and training needed to understand if, why, or how something has happened within the volume and complexity of the Big Data. Ultimately in a business context this understanding has to be converted to a replicable, operational model that can be implemented in a production environment, a task that often is beyond the training and scope of actuaries. While data storage has allowed the accumulation of volumes of data, making sense of it and answering business questions require that we develop the types of algorithms to organize granular data into something that is reportable, understandable, and, more than anything, replicable. As an example: in the 1990s modellers developed algorithms to group together the 15,000 ICD-9-CM medical diagnosis codes into more manageable and useful condition categories. These categories have become the basis for the practice of risk adjustment (see Duncan, 2011), which is the basis for the reimbursement of health insurers in a number of countries. But we lack this type of algorithm for much of the clinical data that is generated, let alone more recent, unstructured social and epidemiological data.

At the same time, an analyst cannot ignore the complexities of the data: how it was generated, how it is coded, what types of coding errors and missing values are included, and how to address any data problems. Structured data generates many complexities and problems of interpretation; these are multiplied with unstructured data that is less well understood. Understanding the data itself, and its sources and limitations, remains critical in understanding the outputs of any modelling exercise.

Big Data analysis: defining the objectives

The needs of running an insurance business have not changed: actuaries need to analyze, select, and manage risks, price appropriately, and ensure the solvency and profitability of the enterprise. The way that actuaries approach problems has changed significantly over the last two generations: as recently as the 1970s, even large insurance companies did not have computers, and policy information was stored on hand-written cards. Actuaries used numerical analysis and mathematical techniques that were developed by Russian mathematicians who did not have computers, as well as other methods that are no longer used (e.g., commutation columns) because of the advance of computer power. Today no one can imagine an insurance company that can operate without significant computer power, the internet, and networking.

While the objectives may remain the same as they were traditionally, the way that actuaries approach them will continue to change. We can expect that Big Data will have a similar effect in the coming years. In the past, statistical techniques were developed to deal with small and structured samples (often drawn, in the case of medical data, in highly-controlled clinical trials). However, the large volumes of Big Data, often collected in an uncontrolled and unstructured way, require new statistical techniques and other methods, for data mining to identify patterns in the "sea of data" and to analyze them (often with non-mathematical and non-distributional techniques) in order to arrive at useful conclusions and predictions. For example, traditional statistics relied on known probability distributions for analysis and modelling. Thus, residuals in a dataset were analyzed to determine the underlying distribution; often a Normal-type distribution was selected due to the ease of computation (even if more rigorous analysis would have suggested another, more complex distribution). Now, with very large volumes of Big Data, do we still need to identify an underlying distribution, or is it sufficient to look for patterns in the data? Clearly, if the goal is to understand underlying patterns, it is sufficient to determine data patterns. But even if a model is to be implemented in a production environment, simulation and scenario processing (non-distributional) techniques may often suffice, as long as they provide good decision rules and enable interpretation of the model to business users.

The traditional statistical approach can be (simplistically) described as follows: propose a hypothesis; search for data; test the hypothesis and, depending on the results, refine it or draw conclusions. The Big Data learning approach, on the other hand, reverses the process: obtain data; spin through it looking for patterns and relationships; develop a hypothesis to explain the findings; test it on additional data; depending on the results, spin again through the data, refine the hypothesis, or draw conclusions.

Big Data, in contrast to traditional statistics, often uses data mining and artificial intelligence-based algorithms such as neural networks and neighbourhood-based clustering algorithms; however, while providing good results these are often hard to interpret and explain to people, due to their complexity and the re-iterated use of the data through the algorithms. There is also the ever-present danger of over-fitting. Caution should be used in applying Big Data techniques—in the words of Emmanuel Candès of Stanford University (2017), "*The big data era has created a new scientific paradigm: collect data first, ask questions later . . . [I]nferences are likely to be false [and] follow-up studies are likely not to be able to reproduce earlier reported findings or discoveries.*"

Professional actuarial standards require actuaries to stipulate their assumptions and their implications. In the traditional statistical approach, this was relatively easily done. But in the Big Data environment, it becomes a much harder task, as often the determination of the patterns and relationships is done by obscure machine algorithms, and the quantity and obscurity of the source data may be so complex that it may be hard to grasp and explain these patterns, thus making the use of the computerized outcomes potentially dangerous. Thus, actuaries have to develop and enhance their “comfort diagnostics”, so as to better apply “sense checking”, get comfortable with the outcomes, and be able to explain them to the public.

The ultimate use of the model or algorithm, and how it fits into the organization’s workflow, is a key consideration. For example, what changes need to be made to the company’s data warehouse to accommodate the data requirements for the application of the model? What is the frequency of data refreshes and how frequently will the model be run? What training will the users of the model results require? Will the model or algorithm be applied for automated real-time decision-making in the operational environment of the organization? How will the model be deployed to end-users, and what changes will it require to their workflow? Too many successful modelling exercises fail because the needs and reactions of end-users are not taken into consideration, particularly when a model automates a process that formerly was the province of a trained professional.

Data quality

Anyone who has worked with structured healthcare data knows that the organization, understanding, and warehousing of data can take longer than the actual analysis. Problems that are encountered include:

- Data completeness: this may arise because of missing observations, or because data is only available on a subset of a population. Techniques exist to complete missing observations; or, if adequate volumes of data are otherwise available, incomplete records may be omitted.
- Data availability on a subset of a population creates a different type of problem, particularly data bias, which can be a significant issue in terms of accuracy of conclusions. Advanced statistical techniques may be helpful for dealing with bias.
- Reporting bias: this problem is particularly important in datasets where there is little or no control over the quality or accuracy of data entry (e.g., social media).
- Lack of standardization and interpretation: because, in the past, we have had at our disposal highly standardized claims datasets, actuaries have not had to deal with the reporting and interpretation problems of, for example, survey data. The concept of “validation” of a survey tool is not familiar to most actuaries, but will probably need to become part of the actuarial toolkit at some point.
- Data aggregation and anonymization: the increasing demand for data privacy makes it difficult to assemble complete datasets for analysis, or to link different datasets with a common identifier. Sometimes data is only available on an aggregate basis that may require different analytical techniques and tools.
- Agreement of the data with statistically known distributions and models: in a structured data set drawn from a known population, with a limited number of variables, it is often possible to

> surmise or determine an underlying distribution and test the agreement of the data with that hypothesis. In a Big Data environment, often drawn from many sub-populations with multiple variables (of which some are selected by obscure computerized algorithms), this assurance of statistical validity is often missing, and various heuristic approximations and simulations are used instead.

The type of problem that can arise when data is inappropriately used or interpreted is illustrated by the Google flu example discussed below. Actuaries should be alert to the problems that can occur in data and be prepared to assemble the necessary resources to address them.

Actuaries' role in Big Data

The ability to analyze and interpret structured and unstructured data requires an understanding of the underlying business and the specific business problems, as well as advanced analytical, statistical, and programming skills. Big Data, with its unstructured, complex, large-volume data, has the possibility and promise of completely changing the way companies do business and data analysts (including actuaries) perform their jobs. At the same time, it complicates the analysis and interpretation of the available data, and requires the skills of data scientists (Conway, 2013); actuaries are well advised to master data science techniques and work in cooperation with data scientists. Actuaries also need to identify the importance and value of Big Data within their organizations, and invest in the appropriate technological infrastructure, analytical tools, and skills. For credentialed actuaries this may require investment in retraining outside their current job functions.

Actuaries have a grounding in statistics and their application in the analysis and evaluation of insurance and other financial risks. Compared with statisticians, actuaries have only a basic level of training in statistics. In part, this is because the actuarial examinations place heavy emphasis on *risk*, and demand deep knowledge of the insurance and financial services environment. The combined knowledge of risk, insurance business processes, and the data that they generate determines the actuarial role in modelling and data analysis in insurance.

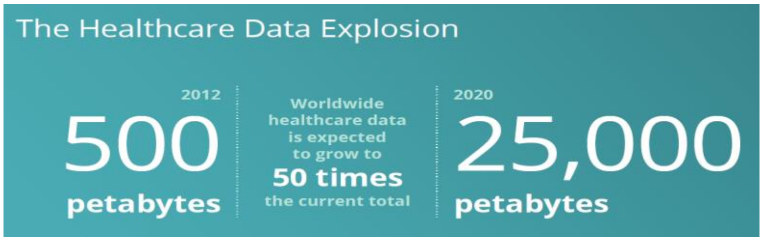
In addition to using Big Data directly in their actuarial work, actuaries may fulfill different roles in a multi-disciplinary Big Data team, from manager (organizing the different disciplines or promoting the business case for a particular solution) and business expert (informing the team of the insurance and risk environment and its particular needs) to data scientist (developing analytics and running and evaluating models). As discussed earlier, a critical need for any model and application is the ability to explain it to the business users; actuaries, with feet in both the business and statistical camps, are ideal for this role.

To enter into and compete in the world of Big Data, actuaries will require a broad range of new skills: new programming and non-traditional analytical skills and techniques; advanced statistical techniques such as regression, generalized linear models (GLM) and time-series; simulation and decision-making techniques; data mining and non-statistical clustering techniques; and artificial intelligence and machine-learning algorithms. Actuaries generally, and particularly those that work in multi-disciplinary teams where their domain knowledge can be applied, will need to be familiar with the more advanced data science tools (even if they are not responsible for applying them). Credentialed actuaries will be required to develop these skills themselves, or be familiar with the tools and their applications, while newly-credentialed actuaries will be required to acquire advanced training in statistics and analytical methods as part of the examination systems. To enhance the skills and knowledge of actuaries in face of the data analysis requirements of Big Data, the two large North American organizations, the Society of Actuaries (SOA) and the Casualty Actuarial Society, have recently changed their qualification requirements to include a broader training in advanced statistics, and (in the case of the SOA) a practical examination at the Associateship level in the application of predictive analytics.[1] The German actuarial association (Deutsche Aktuarvereinigung) is planning a CERA-like qualification in data science.

Either way, some familiarity with the power of new data-handling technologies (particularly in respect of unstructured data) will help actuaries to understand and identify the opportunities that Big Data provides.

Why is Big Data particularly relevant to healthcare actuaries?

Actuaries within the healthcare industry have access to many potential sources of data that could provide insight into risks and opportunities, much of which was not available before. In addition to traditional claims, morbidity, and demographic data, these sources include data generated by fitness devices, wellness devices, medical equipment (including diagnostic devices), and social media. Additional data may be obtained for epidemiological and public health studies from various regulatory and bureaucratic sources in other governmental agencies. The data may be generated by policyholders, patients, health providers (e.g., doctor notes written on an electronic health record), or by diagnostic or other medical equipment (e.g., X-rays, MRIs, or blood test results). Some sources of data did not exist before, such as the mapped genomes of patients, in the context of personalized medicine. This data can have a variety of applications in health insurance and public health, but of course also raises many questions about the way in which insights flowing from it are applied, and the risks posed by its existence. The amount of available data is growing at an astronomical rate, as demonstrated by the following summary:


The Healthcare Data Explosion
2012
500
petabytes
Worldwide healthcare data is expected to grow to
50 times
the current total
2020
25,000
petabytes

(Feldman, Martin, & Skotnes, 2012)

Healthcare actuaries are closely involved in the management of healthcare risks. Historically, healthcare actuaries have managed this risk through of a combination of underwriting, pricing, benefit design, and contracting with providers. However, through the use of Big Data, they are starting to develop unique

[1] The IAA Big Data Working Group is assembling a compendium of data science professional qualifications and continuing-education initiatives in multiple countries.

insights into how behavioural factors affect healthcare outcomes. For example, the success rate of a particular treatment may be dependent on the genetic profile of a patient and his level of fitness. The personalization of medicine[2] requires new data to be entered into electronic health records, with the aim of choosing far more appropriate and even personalized treatment for individual patients, and hence potentially significantly improving health outcomes, and therefore also population mortality and morbidity. For example, knowledge of an individual's genome allows doctors to better match the most effective cancer drug with the individual patient (Garman, Nevins, & Potti, 2007). This may lead to considerable savings in the healthcare industry and reduce wastage on incorrect treatment.

In jurisdictions in which it is legal to do so, Big Data can be used to price and underwrite insurance, and identify targets for marketing. Big Data aids in identifying risks such as medical conditions that otherwise may not be disclosed through the underwriting process, and forms the basis of a rate more appropriate to the risk presented by the insured. In jurisdictions in which this type of underwriting is not permitted, use of Big Data-based risk identification may still be useful for identifying potential high-cost members and candidates for intervention programs, such as wellness, case, or disease management programs that are operated by a number of insurers.

A number of companies are also springing up to capture real-time or near-real-time data from devices such as continuous glucose monitors, scales, pedometers, fitness monitors, and smart watches. With appropriate statistical analysis to identify the range of normal readings, these companies have the ability to target interventions to patients in a more timely or focused way. As indicated before, there is still the signal/noise problem, and more work is often necessary to identify ranges of normal readings under highly variable conditions in order to avoid false positives or false negatives.

At the same time, these devices and their use can be viewed as intrusive by some patients, particularly when they are sponsored by risk-takers or insurers. The tension between those that represent risk (patients) and those that accept it (insurers or medical groups) will continue to be a challenge, as the latter take advantage of new tools to acquire more detailed data on patient risk, and patients and consumer advocates continue to push back on intrusion of privacy.

In some environments, health insurers are the custodians of electronic health records. To the extent that the information mentioned above enters the record, it would, in theory, be available to health insurers. If this is the case, it could be applied in very effective ways to make relevant information available to treating doctors, and hence improve health outcomes. On the other hand, such data is of course very sensitive, and privacy and ownership considerations are very important.

However, to the extent that new sources of medical data are not available to insurers, either because they are legally prevented from requesting it or, even if they ask for it, it is withheld by potential policyholders or providers, there are clear risks of adverse selection in purchasing health or life insurance. In some jurisdictions, it is not clear that insurers would have any rights to access genetic

[2] IAA Health Committee paper: www.actuaries.org/LIBRARY/Papers/HC_Personalised_Medicine_Paper_Final.pdf

information, or other health record information that may be relevant to underwriting, and this may create significant risk.

It is also relevant that much of this data can be used to drive behaviour change in the interest of better health outcomes. For instance, capturing more data on clinical outcomes and augmenting it with geo-location data of the insured and provider allows for high-quality provider networks to be created, and insured patients may be incentivized or directed to use healthcare providers who provide higher-quality treatment. At a member level, any data on wellness activities (whether in the form of preventative screenings, exercise, or nutrition) may be used to incentivize and reward wellness engagement, which in turn reduces healthcare costs for those that respond to such incentives. Determining the optimum level of rewards and wellness activity is an actuarial problem which can be solved if multiple sources of wellness and health data are shared with an insurer.

Text mining doctors' notes on claims or health records can also provide additional information, over and above the procedure and International Classification of Diseases (ICD) codes that would typically be obtained from the claim; this type of data is a "classical" example of a Big Data source, with all the characteristics noted above. This will provide additional information on the complexity of the procedure and the stage of the disease, which will assist in analyzing the success rate of treatment provided. It may also be used to determine the case mix of patients visiting a provider, which may be used in the context of provider profiling, and which in turn gives insights into quality and efficiency of treatments provided.

Big Data can also be used to provide insight into the incidence and spread of disease within a population, perhaps even before individuals access healthcare facilities. For example, Google has used the number and type of searches to produce current estimates of flu and dengue fever in a particular area (Google Flu Trends, 2016), although with varying rates of success. The initial model built by Google failed to account for shifts in people's search behaviour and therefore became a poor predictor over time. Further work has been done by Samuel Kou that allows the model to self-correct for changes in how people search, and this has led to more accurate results (Mole, 2015). This data can provide an understanding of the spread of disease within a population, which can potentially be used as an early warning to identify a potential increase in claims and demand for healthcare resources before it occurs. The Google experience is, however, a cautionary tale about what can happen when machine learning is applied to data without knowledge or understanding of the data or the underlying process by which it is created. For an objective discussion of shortcomings of the potential and problems of artificial intelligence solutions in healthcare see "*Is IBM Watson a 'Joke'?*" (Bloomberg, 2017).

Healthcare actuaries have unique domain knowledge, which means that they are in a position to practically apply these non-traditional data sources to enhance their understanding, solve problems, and seek opportunities. Big Data has the potential to enhance the healthcare industry, through enabling wellness programs to operate effectively, personalizing treatments, and improving the allocation of healthcare resources to reduce wastage in the system. Actuaries also understand financial risk, which is critical to finding the correct application of Big Data tools in insurance.

There are many concerns about privacy, data security, and the ways in which data is used that must be addressed before data is applied in practice. Patient and doctor permission, depersonalization of data for analytical purposes, fail-safe access control to sensitive data, and an ethics and governance framework for evaluating the application of insights to practical problems, must all be in place. Health actuaries need to evaluate the regulatory requirements and the ethics of Big Data applications.

A number of jurisdictions have begun to regulate the use of Big Data. Some actuarial organizations have published actuarial standards with respect to data (for example the U.S. Actuarial Standards Board's standard no. 23: Data Quality). In the UK, the Financial Conduct Authority (FCA) has published a call for inputs on the use of Big Data.

> *In Europe, the European Commission is placing a big bet on big data in its strategy for economic growth. The EU's 2015 Digital Single Market Strategy targets big data as 'central to the EU's competitiveness' and a 'catalyst for economic growth, innovation and digitisation across all economic sectors [. . .] and for society as a whole.' The European Commission published a General Data Protection Regulation in April 2016, regulating (among other things) data privacy and "profiling"*[3]

(Kelly, Blythe, & Long, 2016)

In the United States there is a "constellation" of laws and regulations governing the use of Big Data, at the federal, state, and local level. On November 16, 2017, France's Institut des Actuaires adopted a professional norm named NPA5, which defines actuarial best practices for actuaries when they use Big Data (personalized data and health data), and the American Academy of Actuaries has established a task force to develop a practice note for actuaries working with Big Data in all practice areas that includes a summary of different regulations.

At the same time, actuaries should also consider the risk implications of their organizations not having access to data that exists, and how these risks can be managed.

[3] Ironically, insurance pricing is all about profiling. The implications of "anti-profiling" regulation have yet to be felt.

References

Bloomberg, J. (2017. July 2). Is IBM Watson a 'Joke?' *Forbes*. www.forbes.com/sites/jasonbloomberg/2017/07/02/is-ibm-watson-a-joke/#3f6e3fa9da20

Candès, E. (2017, August). "What's Happening in Selective Inference?" The 2017 Wald Lectures, Joint Statistical Meetings, Baltimore. https://statweb.stanford.edu/~candes/talks/Wald1.pdf

Code Project. (2009, April 19). Distributed and parallel processing using WCF. Code Project. www.codeproject.com/Articles/35671/Distributed-and-Parallel-Processing-using-WCF

Conway, D. (2013, March 26). The data science Venn diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Duncan, I.G. (2011). *Healthcare Risk Adjustment and Predictive Modeling*. Actex Publications.

European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=ENta), and repealing Directive 95/46/EC (General Data Protection Regulation).

FCA. (2016, September 21). FS16/5: Call for inputs on Big Data in retail general insurance. www.fca.org.uk/publications/feedback-statements/fs16-5-call-inputs-big-data-retail-general-insurance

FCA. (2016, September 21). Occasional Paper No. 22: Price discrimination and cross-subsidy in financial services. www.fca.org.uk/publications/occasional-papers/occasional-paper-no-22-price-discrimination-and-cross-subsidy

FCA. (2016, November 22). The challenges for insurance and regulators in a Big Data world. www.fca.org.uk/news/speeches/challenges-insurance-regulators-big-data-world

Feldman, B., Martin, E., & Skotnes, T. (2012). Big Data in healthcare: hype and hope. Dr. Bonnie 360°. www.scribd.com/document/107279699/Big-Data-in-Healthcare-Hype-and-Hope

Garman, K., Nevins, J., & Potti, A. (2007). Genomic strategies for personalized cancer therapy. *Human Molecular Genetics*, *16*(R2): 226–232.

Gartner. (2016, October 18). Big Data. www.gartner.com/it-glossary/big-data

Google Flu Trends. (2016, October 18). Google flu trends data. www.google.org/flutrends/about

Kelly, C., Blythe, F., and Long W. (2016, October 24). How big will big data be under GDPR? The Privacy Advisor. https://iapp.org/news/a/how-big-will-big-data-be-under-the-gdpr

King, G. (2016). Preface: Big Data is not about the data! http://gking.harvard.edu/files/gking/files/prefaceorbigdataisnotaboutthedata_1.pdf

Mole, B. (2015, September 11). New flu tracker uses Google search data better than Google. Ars Technica. http://arstechnica.com/science/2015/11/new-flu-tracker-uses-google-search-data-better-than-google

SAS. (2016, October 25). Machine learning: what it is and why it matters. www.sas.com/en_us/insights/analytics/machine-learning.html

Schneider, C. (2016, May 25). The biggest data challenges that you might not even know you have. IBM Watson. www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know