Alessandro Di Bucchianico
Department of Mathematics and Computer Science
Eindhoven University of Technology
a.d.bucchianico@tue.nl

Laura Iapichino
Department of Mathematics and Computer Science
Eindhoven University of Technology
l.iapichino@tue.nl

Nelly Litvak
Department of Applied Mathematics
University of Twente, and
Department of Mathematics and Computer Science
Eindhoven University of Technology
n.litvak@utwente.nl

Frank van der Meulen
Department of Applied Mathematics
Delft University of Technology
f.h.vandermeulen@tudelft.nl

Ron Wehrens
Biometris
Wageningen University & Research
ron.wehrens@wur.nl

Research

Mathematics for Big Data

This essay highlights several examples of using mathematics and statistics to analyse problems involving Big Data. More often than not, mathematics is essential for extracting usable information from the data. However, it usually remains hidden under the bonnet, and the general public seems to take it for granted. With this paper Alessandro Di Bucchianico, Laura Iapichino, Nelly Litvak, Frank van der Meulen and Ron Wehrens want to show some essential contributions of the fields of mathematics to Big Data using successful real-life examples.

'Big Data' has become a buzz word in the last decade, both in science and among the general public. Scientists from all areas encounter this in the shift of content and methods in their research as well as in current scientific funding programmes. For example, Big Data is one of the selected routes in the Dutch National Scientific Agenda (NWA) and the large funding programme Commit2Data has been launched in the Dutch Digital Delta in 2016.

As the Big Data Team of the 4TU Applied Mathematics Institute, we feel that mathematicians should actively engage in Big Data activities. It is the goal of this paper to show the importance of mathematics in Big Data.

The role of mathematics is easy to overlook and not fully recognized because technological advances are much more visible than mathematical advances even though the latter often have more impact. Here is a small illustration. It is common knowledge that the speed-up of computers due to technological advances follows Moore's Law: doubling of speed every eighteen months. However, it is much less known that the speed-up due to advances in mathematical methods in scientific computing and optimization is at least of the same order of magnitude, and in some areas even much higher (see [6, 23]).

In this essay we present several explicit real-life examples of the mathematics behind Big Data, highlighting the role and importance of specific areas of mathematics in these contexts. We show a wide variety of examples: search engines, virtual prototyping in manufacturing, data assimilation, web data analytics, healthcare, recommendation systems, genomics and other omics sciences, and precision farming. In this way, we hope to stimulate mathematicians to work on topics related to Big Data, as well as to encourage industries and researchers in computer science and other fields to collaborate with mathematicians in this direction.

Similar and more detailed accounts have appeared at other places, see, e.g., [11, 19], National Research Council (2013) and the London Workshop Report on Statistics and Science (http://bit.ly/londonreport).

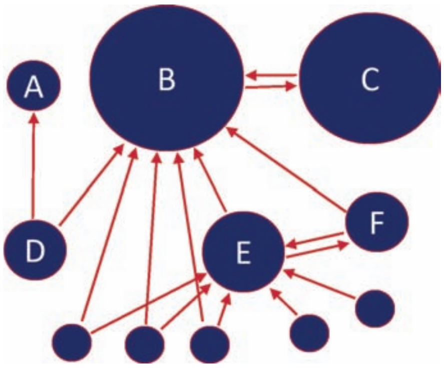

Figure 1 PageRank, example from Wikipedia.

Search engines

The quality of a search engine depends greatly on ranking algorithms that define in which order web pages will appear for the user. This is indeed crucial because most of us do not go beyond the first page of search results. Google's PageRank, at the very heart of the success of Google, was the first and most famous ranking algorithm.

The revolutionary idea of Google was that the importance of a web page depends on quantity, but also on quality of links that point to this page. This can be seen on a small example from Wikipedia in Figure 1.

The size of the nodes represents their PageRank score. Node B has a large PageRank because it has many incoming links. The PageRank of node C is high because it received the only outgoing link from the important node B. Mathematically, the World Wide Web is modelled as a graph with pages as nodes and hyperlinks as directed edges, and then a large set of equations is solved to find the PageRank values for each node in the graph.

Right after PageRank was introduced, its fast computation became a problem of great interest because the Google matrix is huge, at the moment it would have hundreds of billions of rows and columns. In the beginning of this century, major speed gains were achieved due to sophisticated new methods from, mainly, linear algebra [5]. Another interesting mathematical and practical problem is the vulnerability of PageRank to deliberate manipulations, such as link farms created intentionally to boost the PageRank.

If we want to predict effectiveness of ranking, it is also important to understand its relation to the network structure. Can we predict the largest PageRank, investigate its stability, pick up a signal from hidden communities? Can we use ranking to detect important changes in the network structure? A lot of empirical results are available but they do not answer these questions in sufficient generality. To solve these and other problems we need to develop new approaches in probability theory and the theory of random graphs (see e.g. [9]).

Pagerank

'Easily bored' surfer. Consider a simple model of a surfer browsing web pages. With probability α, the surfer follows a randomly chosen outgoing link of a page, and with probability $1-\alpha$ the surfer is bored and jumps to a random page. Initially, Google used $\alpha=0.85$. PageRank of a page is the stationary (long-run) probability that the surfer visits this page.

Eigenvector. Equivalently, PageRank is the so-called dominant left eigenvector of the transition matrix of the above process: the entry (i,j) of this matrix is the probability that the surfer on page i will proceed to page j. Such an eigenvector is unique. The PageRank of a web page is the corresponding component of this unique dominant left eigenvector.

Virtual prototyping in manufacturing

High development costs in industry have led many manufacturers to replace building and testing physical prototypes by virtual prototyping, i.e., testing using large-scale simulations of extensive mathematical models based on physical principles. Specific examples are the automotive industry and the aircraft industry (see, e.g., the Virtual Hybrid Testing Framework of Airbus). Such simulations should be handled with care since there is uncertainty in the outcomes due to both model limitations and the numerical accuracy of the simulations, often requiring solving large systems of differential equations. On the one hand there is uncertainty due to replacing physical reality by a mathematical model. This involves both the uncertainty caused by the modeling simplifications (structural uncertainty) and the uncertainty in knowing model parameters (parameter uncertainty). On the other hand, given a complicated mathematical model, it is important to know how accurately numerical methods can approximate specified outputs from this model.

The term Uncertainty Quantification is often used as general term for scientific research in this area. There exist several mathematical approaches to study this uncertainty. One such approach is applying statistical techniques related to experimental design for computer experiments like Latin hypercube sampling and response surface methods. Another approach is to cast the mathematical model as a stochastic partial differential equation and try to solve that. Recent high-level mathematics combining analysis and stochastics is used such as perturbation expansion methods for random fields, stochastic operator expansions and polynomial chaos (Wiener chaos).

Model order reduction (MOR) techniques (see, e.g., [23]) have been recently introduced and exploited to overcome the issue of severe computational times required for solving mathematical models of real-life processes. Over the past four decades, reduced-order models have been developed aimed at replacing the original large-dimension numerical problem (typically called high-fidelity approximation) by a reduced problem of substantially smaller dimension. Depending on the context, there are different strategies to generate the reduced problem from the high-fidelity one, e.g., Krylov subspace based methods, moment matching techniques, proper orthogonal decomposition, balanced truncation, reduced basis methods. Very short CPU times and limited storage capacities demanded today by MOR methods allow to tackle a wide range of problems arising in engineering, computational science, and physical and biological sciences.

Data assimilation

Weather forecasting, for some people the main reason to watch the news, is a data-intensive computational problem with many economic implications (agriculture, hospitality business, airlines, healthcare, large public events). The change over time of measurable atmospheric quantities can be described in terms of dynamical systems, transferring information in time-ordered observed data to a physical model of the system. This process is often referred to as data-assimilation. Its development has been highly influenced by professionals working in the atmospheric and oceanographic sciences. When discretized in space, a typical model for numerical weather prediction is a differential equation system with dimension of order 10^9 [18]. The state variable of the dynamical system may represent unknown quantities such as for example velocity, temperature and pressure at a grid of locations.

The application of mathematical models to large dynamic data sets has naturally popped up in many other communities as well. Within signal processing recovering the unknown state of the dynamical system is known as filtering or smoothing, where the first term refers to online recovery (as opposed to static recovery). Probabilists and statisticians usually speak of state and parameter estimation. Over the past thirty years there has been tremendous progress for this kind of problems. Under specific assumptions on the dynamical system computationally efficient methods such as the (ensemble) Kalman filter can be used. In more general settings, a Bayesian formulation of the problem and application of Markov Chain Monte Carlo methods and Sequential Monte Carlo methods can be exploited (see, e.g., [21,22]). Whereas these methods are presently not yet applicable to weather forecasting, they have proved to be powerful in simplified problems of less demanding dimensions and constitute a very active area of research [8,17].

Web data analytics

Many companies collect large amounts of customer data through their web services. However, having these data does not mean that we already know everything. Even simple tasks like counting the number of distinct records in a large customer database (e.g., the number of distinct customers that use a certain service) requires advanced mathematics. The exact counting is computationally prohibitive mainly because we cannot keep all objects in the restricted working memory of a computer. However, we might not need that level of accuracy — in such cases it is often sufficient to work with approximate estimates.

HyperLogLog

Hash functions. Each digital object is converted to a sequence of zero's and one's using hash functions. On a set of different objects a good hash-function appears as if randomly generated: zero's and one's have probability $\frac{1}{2}$, independently of each other.
Count zero's. The idea of LogLog-type algorithms is to sweep through objects keeping in the memory *only the largest number of zero's* at the beginning hash functions. For example, if we observed

$$00101,\ 10011,\ 01010,$$

we will remember 2, the largest number of zeros. Roughly, the probability to see 2 zero's followed by one at the beginning of the hash function is

$$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8},$$

so we conclude that we saw approximately 8 objects!
HyperLogLog. In this form, the estimation is obviously *too rough*, so it cannot be directly used in practice. A lot of mathematics went into making the result more precise. This includes dividing hash functions into registers, using different corrections for small and large samples, harmonic averages. All these ideas are included in HyperLogLog, ensuring its applicability. Further improvements are possible, e.g., this was the goal of the paper [15].
Why LogLog? Assume we have N objects. Then hash functions have length $\log_2(N)$. Hence, the number of zero's is a number between 0 and $\log_2(N)$, so we need only $\log_2\log_2(N)$ bits of memory to remember this number.

Probability theory has been essential in developing algorithms such as Count-Min Sketch, MinHash and HyperLogLog that use random hash functions to store answers. Such algorithms may be accurate within 2% while using only memory in the order of the (iterated) logarithm of the original sample size. An important issue in developing in these algorithms is to control the variance of the estimators, in order to get consistently accurate estimates.

HyperLogLog is one of the most elegant mathematical solutions for counting distinct objects in Big Data applications, widely used in practice. Researchers at Google [15] state that Google's data analysis system PowerDrill routinely performs about five million 'count distinct objects' computations per day. In about one hundred cases, the resulting number is greater than one billion. In 2014 HyperLogLog was implemented by Amazon's data structure store Redis as well. An interesting human interest note: the commands of HyperLogLog begin with PF - the initials of the French mathematician Philippe Flajolet who developed this algorithm (see, e.g., [12]).

Maybe even more exciting from a scientific point of view was the result in [3] where HyperLogLog was used to accomplish an incredible task of computing average distances in the complete Facebook graph of more than 700 million nodes. It turned out that the distance (the number of hops along the edges of the Facebook graph) between two Facebook users is on average less than 4!

Healthcare

Medical devices like MRI scanners obtain large image data at relatively low velocity. Efforts are undertaken to reduce the time it takes to makes scans (typically thirty minutes) since hospitals could obtain higher efficiency of the expensive MRI equipment and patients would suffer less from the unpleasant high noise levels. Making scans at a lower resolution is not an option because of medical reasons. An MRI scan uses magnetic fields to order the spins of hydrogen atoms and radio waves to disturb these spins. When the spins return to their original position, energy is emitted. This energy is measured so that one gets an indication of the amount of tissue. Using magnetic gradients it is possible to localize these measurements.

The mathematical bottom line of this procedure is that MRI scans produce Fourier coefficients one by one. Traditional approaches to reconstruction algorithms cannot yield the desired reduction of scanning time because of the so-called Nyquist–Shannon criterion. Again, advanced mathematical techniques have provided the breakthrough. The basic idea is to project the observed data onto a smaller subspace using sparsity in the data. Remarkably, random projections yield sampling strategies and reconstruction algorithms that outperform traditional signal processing techniques. These methods are known under the name compressed sensing. For other applications of compressed sensing in healthcare, we refer to https://www.healthcare.siemens.nl/magnetic-resonance-imaging/clinical-specialities/compressed-sensing.

Compressed sensing has been applied successfully in a wide range of other tasks as well, including network tomography, electron microscopy, and facial recognition.

Recommender systems

Webshops like Amazon analyze the buying behaviour of their customers and present visitors of the Amazon website with recommendations of books and other items that may be of interest. In a similar way Netflix gives suggestions for movies to its customers. A way to provide such recommendations is to set up a matrix of user ratings of movies (columns are ratings, rows are users). Of course, such a matrix has many empty entries since there are many more movies (Netflix has around 20,000) than people can see and rate.

The idea behind the recommender systems is that there are relatively few 'latent' features that drive our preferences (a sparsity principle). That is, there are a few typical items (books or movies) and a few typical users. Translated into matrices, this means looking for a nonnegative matrix factorization of the preference matrix. This means that a very large and sparse preference matrix is presented as a product of two matrices with much lower dimensions. Although computers become faster, this is mainly increase in CPU speed and much less in faster memory. Factorizations of large matrices, however, require a huge amount of communication between working memory and storage memory. There is thus a need for memory efficient factorization algorithms that go far beyond traditional factorization algorithms for singular value decompositions (see, e.g., [16]) for a technical account by the team that won the One Million Dollar Netflix competition). An exciting new approach in this field is the use of randomized methods like stochastic gradient algorithms (see [1]).

Genomics and other omics sciences

Now that technology has become available (and affordable!) to rapidly obtain information about the genetic composition of biological samples, huge quantities of data are generated routinely. This is not only true when looking at genetic information (hence the term genomics) but also when looking at proteins (proteomics) and metabolites (metabolomics), to name just two other members of the 'omics' family. The Big Data aspect here refers to the huge amount of information that we have on a relatively small number of subjects. A typical example is genetic information on humans, animals or plants that consists of millions of measurements (data points) for each subject. The resulting 'high-dimensional' data require the development of new statistical techniques to draw correct conclusions because traditional statistical methods for such data lead to an unacceptable high number of false positives (see, e.g., [7]).

Furthermore, advanced data processing methods are needed to convert the measured data into information — one example is the BLAST algorithm [2], incidentally also the most highly cited paper of the nineties) to align sequences of nucleotides or amino acids with database entries. In each case we are confronted with the issue mentioned before: we know an awful lot about very few samples, which makes statistical analysis extremely hard. Typical questions are finding genes, proteins or metabolites related to certain traits or treatment effects. Network analysis is getting more and more attention (see, e.g., [20]) as a means to bring experimental results into the realm of the things we already know about the biology of the system — one of the main challenges is to combine the different omics data layers into coherent models that explain the behaviour of the system under study [14].

Precision farming

Agriculture is rapidly becoming a data-rich environment, tractors currently being connected to the Internet 24/7 and resembling computers on (large) wheels rather than the dusty and primitive muscle-machines they were in the 20th century. As a result, new questions can be addressed that were unthinkable only ten, twenty years ago: by combining several different information sources (satellite images, plant growth models, management data on plot level) the farmer can, e.g., try to devise optimal strategies to deliver the right amount of water and nutrients to his land and in this way obtain the highest possible yield (see, e.g., [4, 10] and many others).

Here, the problems are the typical big-data problems: even assuming one has access to all databases and knows how to read and use the data, it is not a trivial question how to combine data with very different characteristics, found in different locations and measured for different purposes. One thing is certain: mathematics and statistics play a pivotal role.

Photo: Shutterstock, MONOPOLY919

Figure 2 Smart agriculture.

Conclusion

Mathematics and statistics, being extremely generic tools, have played an important part in technological and scientific developments over the last centuries, and will continue to do so also in this Big Data era. Not only will they contribute to solving problems faster and more efficiently, they will expand our horizon, exposing questions that we never thought about and maybe did not even expect to be solvable. It is important to realize that advances in this area have both a push and a pull component: without being confronted with real-life problems we might lack the incentive or the direction to pursue promising avenues, but without fundamental knowledge we simply lack the tools to tackle the problems successfully. This was expressed in a concise way by Bin Yu in her 2014 Institute of Mathematical Statistics presidential address:

> “Work on real problems, relevant theory will follow.”

(see http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science). Hence the stress on the applications in this paper: mathematics needs them, just like the applications need mathematics.

References

1 C.C. Aggarwal, *Recommender Systems*, Springer, 2016

2 S. Altschul, W. Gish, W. Miller, E. Myers and D. Lipman, Basic local alignment search tool, *Journal of Molecular Biology* 215(3) (1990), 403–410.

3 L. Backstrom, P. Boldi, M. Rosa, J. Ugander and S. Vigna, Four degrees of separation, *Proceedings of the 4th Annual ACM Web Science Conference*, 2012, pp. 33–42.

4 J. Behmann, A. Mahlein, T. Rumpf, C. Römer and L. Plümer, A review of advanced machine learning methods for the detection of biotic stress in precision crop protection, *Precision Agriculture* 16 (2015), 239–260.

5 P. Berkhin, A Survey on PageRank Computing, *Internet Mathematics* 2(1) (2015), 73–120.

6 R.E. Bixby, A brief history of linear and mixed-integer programming computation, *Documenta Mathematica* (2012), 107–121.

7 P. Bühlmann and S.A. Van De Geer, *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer, 2013.

8 A. Cuzol and E.A. Memin, Stochastic filtering technique for fluid flow velocity fields tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(7) (2009), 1278–1293.

9 N. Chen, N. Litvak and M. Olvera Cravioto, Generalized PageRank on directed configuration networks, *Random Structures & Algorithms* 51(2) (2017), 237–274.

10 D.E. Clay, S.A. Clay and S.A. Bruggeman, eds., *Practical Mathematics for Precision Farming*, ASA, CSSA and SSSA, 2017.

11 J. Fan, F. Han and H. Liu, Challenges of Big Data Analysis, *National Science Review* 1(2) (2014), 293–314.

12 P. Flajolet, E. Fusy, G. Olivier and F. Meunier, Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm, in *AofA'07: Proceedings of the 2007 International Conference on Analysis of Algorithms,* 2007.

13 *Frontiers in Massive Data Analysis*, National Academies Press, 2013.

14 R.D. Hawkins, G.C. Hon and B. Ren, Next-generation genomics: an integrative approach, *Nature Reviews Genetics* 11 (2010), 476–486.

15 S. Heule, M. Nunkesser and A. Hall, HyperLogLog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm, *Proceedings of the 16th International Conference on Extending Database Technology,* 2013, pp. 683–692.

16 Y. Koren, R. Bell and C. Volinsky, Matrix factorization techniques for recommender systems, *Computer* 8 (2009), 42–49.

17 K.J.H. Law and A.M. Stuart, Evaluating data assimilation algorithms, *Monthly Weather Review* 140 (2012), 3757–3782.

18 K.J.H. Law, A.M. Stuart and K.C. Zygalakis, *Data Assimilation. A Mathematical Introduction*, Springer Texts in Applied Mathematics, Vol. 62, Springer, 2015.

19 B.G. Lindsay, J. Kettenring and D.O. Siegmund, A Report on the Future of Statistics,' *Statistical Science* 19(3) (2004), 387–413.

20 K. Mitra, A.R. Carvunis, S.K. Ramesh and T. Ideker, Integrative approaches for finding modular structure in biological networks, *Nat. Rev. Genet.* 14 (2013), 719–732.

21 C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer, 2004, 2nd edition.

22 S. Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.

23 W. Schilders, Introduction to Model Order Reduction, in *Mathematics in Industry Model Order Reduction: Theory, Research Aspects and Applications*, Springer, 2008, pp. 3–32.