SASA Presidential Address 2016

It is my honour to present to you the presidential address at the 58th annual conference of the South African Statistical Association. The photograph on the cover of the conference program and abstract booklet was taken on a morning during the week of the 19th September while we were all making our way by foot up the hill to campus, having parked our cars somewhere near Main Road since protesting students had blocked the entrances to campus. The dark sky and clouds are symbolic of what was to come during the next six weeks when both on our campus and many campuses around the country, students and staff were at odds with one another over fee and cultural issues that brought violence and destruction to these campuses. One consequence of this was that we had to move this conference to an off campus venue and while we are pleased that the River Club could accommodate us, we are also sad that we could not host you on our beautiful campus at the foot of Table Mountain.

However, the rainbow on this cover photograph can be seen as a sign of hope and optimism and we are pleased that we have over 350 delegates attending this conference coming from both academia and industry to share their work and ideas in the areas of Statistics and Data Science. We have an exciting group of international visitors from whom we hope to learn about the latest developments in Statistical Learning for Data Science, Correspondence Analysis, Statistical Ecology, Astrostatistics, Computational Developments in High Dimensional Models and Change Point Analysis.

When trying to decide on a theme for this talk, I recalled our original objective for this conference, which was to bring statisticians and other quantitative researchers from around the country and abroad together in Cape Town to share their research, to discuss emerging ideas and challenges in their respective fields and to showcase our collective achievements in the field of Statistics, Analytics and Data Science. In my mind this conference was meant to be a celebration of our achievements as opposed to a focus on the crisis related to the shortage of statisticians in academia and to some extent

elsewhere in the country. Then the "fees must fall" protests happened, and associated with that the demands for a more relevant curriculum that speaks to Africa and its problems. And so another theme emerged, that of Statistics for Africa.

My celebration of the good things in our world of Statistics will thus focus firstly on how exciting it is to be living in an era where everyone is excited about data and where Data Science is one of the biggest buzz words out there and secondly on how we use our statistical expertise to solve African problems. I want to take you on two journeys today, one a journey through time that will look at the history of Data Science with a specific focus on the role of Statistics in Data Science, and a second virtual journey through our beautiful country that will visit the different universities and research organisations and highlight the exciting research that is happening at these institutions to solve African and South African problems.

We start with the Data Science Journey. What is Data Science? Is it correct to say that it is just another word for Statistics and that we should all just relabel our discipline from Statistics to Data Science? It would certainly help to attract students to our discipline. After all, borrowing from the IMS Presidential address at the 2014 IMS Annual Meeting, when comparing the Wikipedia definition of Statistics

"Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments. When analyzing data, it is possible to use one of two statistics methodologies: descriptive statistics or inferential statistics."

And of a statistician,

"A statistician is someone who works with theoretical or applied statistics. The profession exists in both the private and public sectors. It is common to combine statistical knowledge with expertise in other subjects…Typical work includes collaborating with scientists, providing mathematical modeling, simulations, designing randomized experiments and randomized sampling plans, analyzing experimental or survey results, and forecasting future events (such as sales of a product)."

With that of Data Science ,

"Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science."

And an IBM website description of what data scientists do,

"Data scientists are inquisitive: exploring, asking questions, doing 'what if' analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommendations across an organization's leadership structure."

it is easy to see that we would rather be known as Data Scientists, than Statisticians! But would that be right, or would we be pretending to be something that we are not, or are not fully?

There are various histories of Data Science to be found on the web and one thing that they have in common is that it all started with Tukey when in 1962 he wrote in "The Future of Data Analysis", "*For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt… I have*

come to feel that my central interest is in data analysis… Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science… How vital and how important… is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being 'important but not vital,' although in others there is no doubt but that the computer has been 'vital.'"

So as early as 1962, Tukey emphasizes the importance of data, and refers to "science" and "computing". Let's jump to 1993, not that much did not happen during the intervening years, but it was in 1993 that John Chambers, co-developer of the S language for statistics and data analysis while at Bell Labs, published his essay, "Greater or Lesser Statistics, A Choice for Future Research". In that he said, *"the statistics profession faces a choice in its future research between continuing concentration on traditional topics – based largely on data analysis supported by mathematical statistics – and a broader viewpoint – based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal..."* This was maybe the first warning to our discipline that if we wished to stay relevant, we had to expand what we did.

References to "data science" popped up in various books (for example, Peter Naur's Concise Survey of Computer Methods in 1974) and conference titles (for example, that of the 1996 Conference of the International Federation of Classification Societies in Japan), but shall we credit William Cleveland for defining it in terms of Statistics in 2001 when he published "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics". The expansion he referred to centered largely on closer collaboration between statisticians and computer scientists and he suggested that *"statisticians should look to computing for knowledge today just as data science looked to mathematics in the past".*

Also in 2001, Breiman in his work "Statistical Modeling: The Two Cultures" distinguished between models for prediction *" to be able to predict responses to future input variables"*, a culture that largely ignored underlying data generating mechanism, made no assumptions, allowed for many different predictive algorithms ,i.e., machine learning, and that of stochastic models for inference, *"to infer how nature is associating the response variables to input variables"* based on an assumption that there is a true model that generated the data.

These algorithms that Breiman referred to are often embedded in computer programs aimed at creating intelligent machines that should be able to translate a phrase from one language to another, drive a car, predict the behavior of shoppers, identify a new galaxy, and so more. Thus one forgets that complex statistical thinking often underlies these Artificial Intelligence algorithms and Expert Systems. Yee Whye Teh, Oxford University professor of statistical machine learning argues (Significance, Vol 12(1),2015) that only when these AI programs moved from being based on logic to being based on probabilistic causal reasoning, were they able to work in a world of uncertainty and noisy data. He credits Judea Pearl who in 1988 published "Probabilistic Reasoning in Intelligent

Systems" and who in 2011 won the ACM Turing Award for his contribution to Artificial Intelligence through the development of a calculus for probabilistic and causal reasoning. Pearl's work was based on Bayesian networks, which he described in his 1985 paper as *"directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal dependencies between the linked propositions, and the strengths of those dependencies are quantified by conditional probabilities."* These networks allows for the updating of information and probabilities based on new information (or data) using underlying Bayesian theory.

In 2009, the first edition of "The Elements of Statistical Learning" by Trevor Hastie, Rob Tibshirani and Jerome Friedman was published. In their words, *"vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all; to extract important patterns and trends, and understand "what the data says". We call this learning from data."* It is a text that speaks to the adaptation of existing statistical methods and the introduction of new approaches to cope with the new challenges that arose as a result of the data revolution and the importance of computing. It is most certainly the text that forms the basis of many of the programs in Analytics and Data Science that are being offered by South African universities. Trevor is of course one of the invited guests at this conference. We learnt from him yesterday in his workshop of the statistical models used by data scientists for inference and prediction and we look forward to his plenary address where he will be able to talk with much more authority than myself about what it requires to be a data scientist.

Alongside all these developments in the academic world, it is really the explosion of the availability of data and the recognition by both business and research that data is knowledge that led to the current hype. It is when Davenport associated "analytics" with "winning" in a Harvard Business Review in 2006 and in his book "Competing on Analytics: The New Science of Winning" (2007), when Hal Varian as Google's Chief Economist endeared himself to statisticians for ever to come when he told the McKinsey Quarterly in 2009 that *"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades"*, when Kenneth Cukier wrote in the Economist in 2010 that "*a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data*", that we finally realised we were onto something good. Currently the Careers24 website lists Statistical Analysis and Data Mining as the top in demand skill in South Africa and the second most in demand skill globally (after cloud and distributed computing).

Getting back to the trying to figure out what is data science, the Alan Turing Institute, established in 2015 in the UK with headquarters in the British Library, (and one of many recent institutes focusing on Data Science and Big Data) describes it as follows:

"As the amount of data we generate increases, so does our need to understand and use that data. Data science is the fundamental science behind data analytics, and draws on various existing sciences in response to that need: the mathematical sciences, the computing sciences, the social sciences, software engineering and domain expertise from multiple industries and sectors."

So Data Science is a collaborative science. In particular, the three main components relevant to us is that it is a merger between Statistics, Computing and Domain Expertise.

It is the marriage between Statistics and Computer Science that is one of the main drivers behind changed approaches for our discipline To be a good statistician or data scientist you need to be able to program. Statistical Software development is of course not new. We are all familiar with packages such as BMDP (well some of us remember this at least), SPSS, SAS, S, STATA and R and many of us also program in Python or the younger generation tells me there is now something called "Julia". These days you do not only publish a paper with your methodology and results, you add the dataset and you add the "work flow", often in terms of an R script. This allows other researchers to reproduce your analysis, or tweak it to make it more relevant to their data, or to improve upon it.
It is in this context that a book such as Hadley Wickam's (Chief Scientist at RStudio) "R for Data Science", to appear later in 2016, is eagerly awaited, though there are of course many similar texts.

So we have reached 2016 in my timeline, but what about the future? I read with interest David Donoho's remarks on the "Science" in Data Science in his "50 Years of Data Science", where he referred to "science-wide meta analyses" where we study all the data analyses published on a given topic, "cross-study analysis" where different models are fit on several datasets in order to identify the best performing model, and "cross-workflow analysis" that compares the impact of different methodologies and protocols on the eventual results and conclusions. He concludes by saying that "in future scientific methodology will be evaluated empirically as opposed to mathematical derivations and proofs." Thus Data Science is the science of learning from data.

Where that data comes from brings one to the third component, that of Domain Expertise. It refers to the many areas of application that are currently generating data and in particular big data and the associated interesting problems. We have the chance to talk to and work with people from Astronomy, Medicine, Mining, Engineering, Ecology, Finance and Linguistics, to name a few. To quote Tukey, *"the best thing about being a statistician is that you get to play in everyone's backyard."* This emphasizes the need for collaborative research, the need for encouraging students to study towards double majors, on in the domain field and the other in data science, and even the idea that maybe the statistician should be housed in the department of Astronomy, or Medicine or Biology or Economics.

It also nicely brings me to my second focus area of this talk, that of Statistics for Africa, and how this emphasis on data-driven statistics and the solving of real-world problems open up the opportunity to all of us to make our discipline relevant and hence either to accommodate the requests of the protesting students or to silence their criticism that we teach and practice a "western, euro-centric" discipline. And so I went on my virtual tour around South Africa, to find out in whose backyards the statisticians and data scientists at the different universities and research institutions were playing. I based my information largely on the abstract submissions to this conference and on information

displayed on the institutional websites. So if I misrepresent your endeavours, or omit some important projects of yours, I apologise but at the same time advise you to go and update those websites.

The department of Statistical Sciences at UCT focuses mostly on applied statistics, and have research groups in Finance, Statistical Ecology, and together with statisticians from the Health Sciences, Biomedical Statistics. At this conference you will hear staff and students from UCT talk about mathematical and longitudinal models for the measurement and elimination of malaria, latent class models for classification of HIV viral load profiles, a bio-marker based incidence estimator for HIV, a population based genetics approach to infer selection from longitudinally-sampled HIV-1 haplotypes, semi-parametric mixture models for longitudinal HIV viral load trajectories, competing risk models for the analysis of childhood pneumonia, joint models for TB pericarditis , occupancy models for species range dynamics, population dynamics and stock assessment of geelbek, identification of frogs and crickets using acoustics, dynamic factor analysis and state-space models for semi-annual waterbird counts, use of hidden markov models for analysis of black eagle acceleration data, species distribution modelling of quiver trees, enhance optimisation of investment portfolios, robust portfolio construction and optimisation and the analysis of key performance indicators in rugby union, to name but a few.

The department of Statistics and Actuarial Science at the University of Stellenbosch's publications of the past two years reveal that they practice their trade in education, the finance of emerging markets, social security, Moravian Bells, Dohne Merino lambs, in addition of course to their strength in multivariate, Bayesian and non-parametric statistics. They have just teamed up with global asset management firm Schroders Investment Management to collaborate on research that both expands academic knowledge and enhances the delivery of investment outcomes by investment professionals.

At the University of the Western Cape, Statistics teams up with Population studies in one department. The research interests on their website include, fertility and reproductive health, child and maternal mortality, women empowerment, biostatistics, education, astrostatistics, financial applications in currency markets, poverty and inequality. Conference abstract titles include addressing the deficits in urban statistics across Africa, Bayesian Cox Proportional Hazards modelling and insights from the censored quantile regression model for paediatric and adolescent HIV/AIDS patients on antiretroviral treatment, and the use of survival analysis to model the woman's waiting time to first birth after marriage in Rwanda.

Members of the department of Statistics at NMMU apply their statistics to the areas of finance, sport and among the titles of their conference abstracts one detects an involvement with municipalities in terms of electricity and water provision. These abstract titles include the use of data envelopment analysis (DEA) and the Malmquist productivity index (MPI) to determine the effectiveness of the national benchmarking initiative in achieving its objective of improved efficiency of water service provision during the course of its implementation, a cost frontier approach to an efficiency analysis of electricity distribution by South African municipalities, generalized autoregressive score models applied to financial time series from the JSE, algorithms for dating cycles in financial time series, analysing the impact of climate change on agricultural productivity in South Africa, the use of hidden Markov models to predict observer-confirmed kill sites from GPS lion relocation data, causality between JSE indexes and exchange rates, models for the detection of non-technical electricity losses

in Nelson Mandela Bay Municipality, and using support vector machines, naive Bayes and K-nearest neighbour classification algorithms to develop statistical indices for economic contribution of cultural and creative industries in South Africa.

The Statistics department at the university of Fort Hare has a strong interest in biostatistics and epidemiology. They will be presenting talks and posters on a mathematical model for transmission mechanism of TB/HIV co-infection, the assessment of risk determinants in the regularity of malaria using Cox proportional hazards modelling and an evaluation of disease mapping models.

At Rhodes, Sarah keeps herself busy by analysing the behaviour of honeybees and wasps while Lizanne and Isabelle continue to be Bayesians.

The University of KwaZulu –Natal applies their statistics to medicine, finance and education. Recent publications include Bayesian spatial semi-parametric modelling of HIV variation in Kenya, the use of Rasch modelling to re-evaluate malaria diagnosis test analyses, flexible random effects distributions in disease mapping models and multiple correspondence analysis as a tool for analysis of large health surveys in African settings.

The main fields in which research is undertaken in the Department of Statistics at the University of Johannesburg are Financial Statistics and Industrial Statistics. The applications are primarily in finance and in industrial quality control. Though Lyness is presenting a paper on the use of quantile regression to identify predictors of blood pressure in South Africa.

The publications listed on the Wits website include time series modeling of paleoclimate data, an evaluation of how a single-factor CAPM works in a multi-currency world, principles of using microdiamonds for resource estimation, and co-integration modelling for empirical South American seasonal temperature forecasts.

Some interesting applications of Statistics at the University of Pretoria include joint work with the CSIR in the field of human language analytics, the development and implementation of a geo-spatial predictive model that could be used to understand the rhino poaching problem in the Kruger National park, robust nonlinear mixed effects regression models in tuberculosis research, brain imaging (FMRI images), extreme value modeling of natural disasters as well as statistics in sport. For example they have looked at applying their actuarial skills for the development of an actuarial method based on life assurance valuation techniques for adjusting cricket teams' scores in rain-affected Twenty20 cricket games, numerical and graphical performance measures for evaluating and comparing batsmen, bowlers and all-rounders in cricket, the construction of a system for ranking of rugby teams from multiple leagues and the optimal allocation of swimmers to relay teams.

At the university of Limpopo, they are concerned about the supply of water and electricity as reflected in the conference abstract titles that include the application of spatial statistics of extremes to the extreme flood heights in the lower Limpopo River basin of Mozambique, an application of small area estimation methods in modelling lack of service delivery at ward level in terms of water, sanitation and electricity in South Africa, and modelling average maximum daily temperature in South Africa using r largest order statistics. Sometimes they team up with statisticians from the University of Venda to model heat waves and their impact on electricity demand.

Northwest University is well known for its theoretical research into non-parametric statistics and they of course have a special fondness for the good old bootstrap! However, on the applied front, they have a research unit for Business Mathematics and Informatics that focuses on applied research in the broad areas of financial risk management and telecommunications applications. From their Mafeking campus, we have an abstract submission on forecasting electricity consumption in South Africa using SARIMA models.

At the University of the Free State, a large part of the department's research into stochastic modelling, multivariate statistics, Bayesian statistics and extreme value theory is performed in collaboration with industry partners such as ESKOM and SASOL. In addition they are involved in researching student performance trends and Sean and Stéphanie used Dirichlet regression, Negative Binomial models for inflated zeros with multivariate random effects, using INLA package and censored logNormal regression to analyse the eating habits of bat-eared foxes.

Let's not forget our research organisations, like the HSRC, where the vision of the Research Methodology and Data Centre (RMDC) is to be the preferred data collection, capturing, analyses and digital data repository service provider for the HSRC's research in support of evidence based human and social development in South Africa and the broader region, and the MRC whose research focuses on the ten highest causes of mortality in South Africa and includes TB, HIV, chronic diseases, alcohol and drug abuse, and women's health, and the Data Science for Impact and Decision Enablement (DSIDE) initiative of the CSIR with project themes that include environmental modelling, geospatial modelling, information systems for public institutions, knowledge graph creation and mining, natural language processing application, predictive policing, process optimisation and resource allocation, and StatsSA that assists the government in working consistently towards eradicating poverty and reducing inequality through their provision of statistical information to inform evidence based decisions.

Our students can be proud of us! It really does sound like a veritable feast of statistical methods and applications, all very relevant to making our country a better place for its people – and so we have lots of reason to celebrate.

My journeys have brought me back to Cape Town and to the here and now, maybe not to our preferred venue and maybe not with the view of Table Mountain that we had originally planned. However, I trust that I have wetted your appetite for all the interesting oral and poster presentations that await you during the next three days at this, the 58th annual conference of the South African Statistical Association.

I thank you.

Francesca Little