Proceedings of the International Conference on Industrial Engineering and Operations Management
Pretoria / Johannesburg, South Africa, October 29 – November 1, 2018

Big data in sport analytics: applications and risks

Euodia Vermeulen
Department of Industrial Engineering
University of Pretoria, South Africa
euodiav@gmail.com

Sarma Venkata Yadavalli
Department of Industrial Engineering
University of Pretoria, South Africa
sarma.yadavalli@up.ac.za

Abstract

This paper briefly describes big data generated by physical activity trackers (or wearables) and its application in sport, with a focus on individual sports of running and cycling which are easily accessible to the lifestyle athlete. It describes the potential for paradigm shifts in training monitoring, rehabilitation, talent acquisition and even urban planning that data mining can provide. Furthermore, it highlights some risk concerns pertaining to big data in sport such as user privacy, data accuracy, interpretation of information and athlete autonomy. The paper concludes by contrasting the possible advances in sport and physical activity research with the ethical considerations that might slow down the progress.

Keywords
sport analytics, big data, training monitoring, wearable devices, research ethics

1. Introduction

This paper seeks to address two, perhaps somewhat opposing, avenues in sport analytics: the possible research advances that can be made based on sports tracking data and the ethical considerations that accompany data mining in sport. Sports analytics is the examination and modeling of sporting performance using scientific techniques (Morgulev et al., 2017). It is an emerging discipline which combines a wide domain of specialties from human physiology and kinetics, sport science, big data, data science, data mining, mathematics, and statistical analyses (Passfield et al., 2016). This paper explores some of the recent uses of big data (that are generated by personal fitness trackers or sporting websites) both within sport as well as in other application fields. It describes sports tracking data in terms of the three "V's" of big data and provide an analog for the transition from tracking data to wisdom.

Sports data can be both quantitative and qualitative and can be collected on a large scale from a variety of sources such as biometric data, films or videos, historical medical reports, on-field or on-route positional tracking data, weather and crowd behavioural data to name but a few. Wearable microelectronicmechanical (MEM) systems can collect a range of biometric (physiological, kinematic and kinetic data) and geo-spatial tracking data as athletes physically move through space during sporting activities (Wilkerson et al., 2016). Wearable devices (from hence on

referred to as wearables) include pedometer anklets, chest straps, running or activity watches and smart phones with monitoring applications. These devices enable data to be captured on an enormous scale that falls in the domain of big data, as nearly each positional change an athlete makes can be tracked. The possibilities for research are extending beyond the fixed clinical settings and into the real world.

Big data has, to some extent, reversed the roles of the researcher and the data: often times the data are collected before the research question is asked. Only upon viewing the data, the researcher asks a question from the data and then attempts to find the answer hidden in the patterns in the data. An analyst must therefore know and understand their data in order to ask the right questions and find meaning in the data. Big data can further be characterised in terms of the three V's (Kitchin, 2015):

- Volume: enormous data set sizes measured in terabytes or petabytes;
- Velocity: data is being created and transmitted in near-real time which results in an extremely fast arrival rate.
- Variety: the organisation of the data is diverse and presented as structured, semi-structured and unstructured.

Further the data is all-inclusive as it attempts to capture the whole population or system of interest. The resolution is granular (low-level of detail) and relational, meaning the data contains common fields that permits data sets to join. The data is both flexible in that it can easily add new fields and scalable with rapid expansion in size (Kitchin, 2015). The data is also highly variable in that it changes quickly. However, there is one characteristic that is not often mentioned in the literature that may make the data questionable: its veracity. Big data may display much commotion and noise with its accuracy sometimes drawn into question (Mashooque et al., 2017). However powerful and insightful the advances that big data might bring, its veracity may be its Achilles heel. Data analysts must keep its conformity to accuracy in mind when constructing statistical models in order not to be misled by outliers that has a gravitational pull on the pattern in the data. On the other hand, the outlier itself may be the actual signal of a hidden phenomenon that justifies further exploration, and not the pattern created by the majority of the data. The outliers may be indicative of biasedness in the data set or perhaps an algorithm that does not adjust quick enough to capture the change in the data as it arrives. Outliers might therefore not be an enemy of the analysis, but also an aid to improve algorithms or lead to other, unexpected meaningful discoveries.

The mentioned characteristics of big data are apparent in tracking data captured by wearables. Cortes et al. (2014) reports on the number of workouts and the resulting large data sets generated over five months by users of fitness applications, in particular runners. One month could reach as much as 37 558 648 workouts and a minimum of 16 510 934 workouts. Global Positioning System (GPS) data are generated as a runner moves along his or her route and is sent from the device to the online server as a tuple. The tuple contains, per time stamp, the latitude and longitude, the distance moved, the pace and the altitude. The maximum frequency of GPS tuple data generation and transmission was roughly 25 000 tuples per second with a minimum of 10 000 tuples per second. The estimated number of tuples generated per month ranges between 2.8 and 6.3 billion (Cortes et al., 2014). Kosmidis and Passfield (2015) collected 2.5 million unique data points in their study on elite distance runners. The data were collected during the whole training and/or race session, with unique data points generated for nearly every second of the running activity. The data were diverse and had a very low level of detail, ranging from GPS locations to physiological data such as heart rate. The data was generated in near-real time as the runner moved along their route and could be synced to an online platform during or right after the run. However erroneous readings which presented as outliers and missing data were reported and had to be cleaned or interpolated by the analysts before data interpretation. The outliers included extremely large values for distances covered in a short amount of time, resulting in humanly impossible speeds. These speeds could have had a dramatic effect on their end product if the researchers did test their data and acknowledged the outliers.

This low-level of granularity of the data, the extreme scale of the GPS tuples and the fats arrival rate of the data complicate the analysis to find true meaning and actual value that can provide decision support to athletes, coaches and other users of the data. The data are not validated against known standards within the field and its veracity is uncertain. In order to extract the correct variables, sport domain experts must work alongside skilled data scientists and analysts who are capable to extract the data and build the analytical tools needed to find the patterns in the data. The domain expert will be able to identify outliers more readily than the data scientist. The research question may also only be asked once the data have been visualized and/or aggregated to some extent, which requires the skills of a data scientists and the domain expert to find the question that needs asking and let the data lead them towards finding the answer.

2. The data to wisdom network

"We are drowning in information but starved for knowledge" - *John Naisbitt*

This statement by John Naisbitt in 1982 is evergreen and may be even more applicable today than in 1982. The term "information" may be replaced with "data". The unprecedented volume of sports tracking data being produced by wearables presents a challenge of its own: what data must be used and how to derive meaning from the extracted data. Even if something can be measured, it does not imply value. The optimal solution is to perform minimal testing and measuring that holds maximum return of information (Sands et al., 2017). Domain knowledge will be useful in deciding which variables' data to capture and analyse. However, once extracted and organised, the data will remain a black box of meaningless numbers until mathematics and statistical methods have been applied to aggregate the data into information.

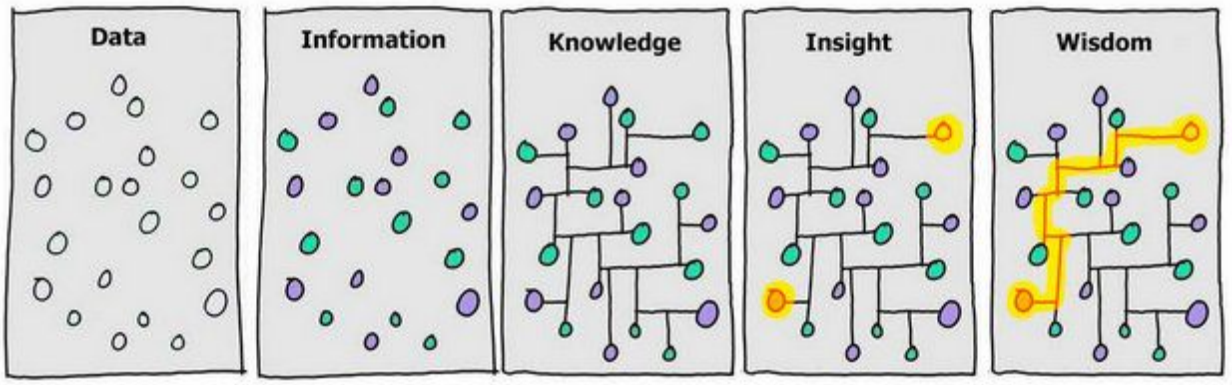

Figure 1: The data to wisdom network. Source: (Kaushik, 2016)

The data-to-wisdom maze cartoon by David Somerville shown in Figure 1 is an analog for the progression of raw data that has been mined from various sources to the beneficial use thereof. Raw data are just random dots on a page with no pattern, i.e. the GPS tuples generated by the device and stored on the server database. Information adds colour to the blank dots and starts to separate the random points. Information provides the athlete with the numbers and figures on how they performed during a run. It tells them how fast they were running, what their step frequency was, how far did they run and much more. However, these will remain just numbers if the runner does not understand them or knows how to interpret them. Knowledge shows the connection between the data points. To gain knowledge, the runner must learn and understand the context of the information and how the metrics are linked. Knowledge allows the runner to start thinking about the information they have at their disposal and how to actually use it to their benefit. Insight provides the starting and endpoints in the maze. When an athlete starts to impart their

knowledge into practice, they develop insight into their capabilities and can use the insight to train better and relay their training into good racing results and general health. Wisdom shows the map to complete the journey from raw meaningless GPS tuples to practical intelligence that provides decision support to athletes and coaches regarding training regimens and race tactics.

3. Research opportunities created by the market for wearables

A quick overview of the market for connected devices and wearables in conjunction with the state of the sport of running and cycling provide a basis to argue the research opportunities created by this union. Cisco Internet Business Solutions Group (IBSG) forecasted that by 2020, there will be 50 billion connected devices in the world that provide applications for healthcare and business services, monitoring of air pollution, transportation, energy use and more (Cortes et al., 2014). The wearable device market increased by 7.3% from quarter three in 2016 to the third quarter in 2017. Globally a total of 26.7 million devices were shipped during the third quarter of 2017 (IDC, 2017). The analysis of public data from the sport tracker service Endomondo by Cortes et al. (2104) covered a total of 333 689 workouts from randomly selected 15 090 users over five months. From this sample, running presented 42.6% of the workouts followed by cycling at 22.87%. Between 1990 and 2013 the United States of America (USA) road race finishers increased by 300%. Running as a sport is valued at $1.4 billion in the USA (RunningUSA, 2016). In South Africa two iconic races have seen much growth since their start. The number of finishers of the Comrades Marathon increased by 37.9% over the last decade. The Old Mutual Two Oceans Marathon started with 26 runners in 1970 and have grown to 26 000 runners in 2017 (OMTOM, 2018) . In South Africa from the month of May 2018 there are 209 cycling events across multiple events left for the year (CyclingSA, 2018). The market size for cycling in the USA was estimated at $6.2 billion in 2015 (Statista, 2015). The opportunity exists for data science and analytics to draw meaning out of the big data generated by a growing running community, a large cycling community and an increasing market for connected devices, including health orientated wearables. The application fields are not limited to physical activity or sport: the geo-spatial metadata from wearables also attracted the attention of urban city planners with the aim of designing infrastructure that support an active lifestyle.

3.1 Quantification of running gait metrics using wearables: from clinical settings into the real world

A wearable device is here defined as a micro-electronic device with on-board sensors that is attached to the human body to study the motions of the athlete during a run activity. Human motion is described in kinematic and dynamic variables pertaining to the anatomic body parts and its displacement, the velocity of displacement and the acceleration and deceleration thereof. These measurements are fundamental to accurately capture human movement through space and time as is required for a sporting activity. The device has an in-built accelerometer that is used to calculate cadence (stride frequency or step rate), stride length and flight time. Furthermore it can measure vertical oscillation and ground contact time (Iervolino et al., 2017). The use of wearables for rehabilitation purposes has the potential to shift the paradigm of the current evaluation methods. The assessment of running mechanics can now be extended to the reality of streets, uneven surfaces, trails, a variety of footwear and responding physiological parameters throughout the run with almost instantaneous feedback. By implementing the use of wearables during running activities, running metrics can be evaluated over a longer period or distance. The assessment of the runner becomes comprehensive and is not subject to the limited spatial and clinical settings nor just the contact time with the medical practitioner (Napier et al., 2017). A single examination or observation does not provide the full picture of the athlete and may lead to wrongful diagnoses of underlying injuries, or missing the injury completely. However, with the wisdom that can be gained from data mining and analysis on a large scale the medical professional will be provided with another tool to evaluate patients with running related injuries. An image of the athlete's running form painted with numbers will provide the practitioner, coach and athlete alike with the opportunity to see a fuller picture and not just isolated evaluations in limited settings.

3.2 Metadata for urban planning for cycling and running

The high resolution of cycling data from the fitness application Strava proved to be a useful supplement in the spatio-temporal analysis of bicycle trip volumes and distances in the Miami-Dade County in the USA. The fitness application data were extracted and used in conjunction with road network characteristics, location specific characteristics and socio-demographic data in a linear regression model to find the bicycle kilometers travelled in the area. The researchers summarised bicycle ridership over three time periods: yearly data stretching from January 2015 to May 2015, weekend data for February 2015 and weekday data for February 2015. The researchers were able to reveal the spatial variation of ridership. The large volume of fitness application data allowed researchers to summarise cycling volumes over a six-month period from January to May 2015 as well as the expected differences in non-commuter and commuter cycling volumes over the week and weekdays. This finding helps city planners to understand cycling behaviour and the environmental impact of the area on cycling volumes for road network planning (Hochmair et al., 2016).

Balaban and Tunçer (2017) analysed data from personal tracking devices to support urban planners to design roads and living spaces to meet the spatial needs of running. The millions of GPS points generated from the runners' devices were used to develop data visualisations of the routes and streets used by runners in Singapore. Other data were included and formed a joint analysis and visualisation of streets' capabilities to accommodate running. This data included the street topology, climate and socio-demographic attributes of the area. The visualisations helped urban planners to understand the behaviour of runners in a specific area and time periods. A total of 38 703 run sessions from 500 000 users were used in the study. The large scale and data from a diverse group of runners gave the researchers a broad and comprehensive overview of the runners' behaviour in the area across seasons and time periods of days. One of the visualistions is an overlay of dynamic timeframes on selected parts of the area where the runners' routes have been plotted. This visualistion identifies patterns in runners' behaviour at different times of the day, for instance streets that are not well lit are used more during the day but not that much at night (Balaban et al., 2017).

In both these studies the metadata from wearables proved useful to answer questions that might nearly be unanswerable using conventional methods such as questionnaires or location based time studies. The research environment was uncontrolled but representative of the real world on a day-to-day basis, capturing weekly and seasonal trends with its inherent variability. By quantifying running and cycling loads from real lifestyle and recreational athletes they were able to open up the research potential to plan for better urban areas that will support an active lifestyle. The tracking data became the eyes and the ears of the researchers on ground level, but on a massive and unprecedented scale that will not be possible without wearables.

3.3 Talent acquisition in cyclists

The career development and success outcomes of elite cyclists were studied by Passfield and Hopker (2016). Coaches and scientists generally accept that elite endurance athletes have been through many years of training to enter and ride in the top level of the sport. However, the progression of success from junior to senior level is still unclear. The researchers completed a retrospective analysis by mining data from race results published by websites. The races occurred between 1980 and 2014. A total of 67 503 race results from 5561 riders who rode in 25 major junior and senior elite races were extracted from the websites using web crawlers. The data analysis revealed the relative age effect, i.e. the Matthew effect in world-class cycling. Cyclists born between January and March dominate the population of riders at World Tour Level. This may be indicative of coaches' bias during identification and development of junior cyclists. A junior cyclist may be overlooked when their birth date is outside of the qualifying range for the age bracket under consideration. The researchers suggested to extend the age bracket with three months on either side, as to include riders born nearer the end of the year.

Cleaning of the data from the websites were reported by the researchers as challenging. Athletes names were misspelt or entities presented with missing variables. The analysis was also eventually limited to the inter-quartile range due to inconsistent findings caused by outliers. This attest to the uncertainty of the veracity of big data. In this case, its inaccuracy may be attributed to human error.

3.4 Training monitoring and prescription in runners

Monitoring of training is not a new concept, but the introduction of wearable tracking devices is shifting the boundaries and the status quo of athlete monitoring. Big data brings with it many opportunities for renewed insight into performance capabilities, on and off the field. Athletic evaluation pertaining to movement and physiological variables are usually done in restricted laboratory settings. The laboratory environment is hardly representative of the real training and racing environment (Iervolino et al., 2017). With on-field monitoring being made possible by wearables which provide constant measurement of important running metrics, the scientific measurement paradigm can be shifted from restricted and re-active laboratory tests to the reality in the field and throughout whole training sessions and races (Passfield et al., 2016). This constant training monitoring of an athlete will provide both athlete and coach the opportunity to identify patterns of response in their biomechanics relating to fatigue, environmental influences such as elevation or uneven terrain as well as physiological responses such as heart rate. The individual's response to training can now be thoroughly analysed, problem areas identified and training adjusted accordingly. Both athlete and coach will gain a better understanding of their inherent capabilities and be able to set realistic, attainable goals.

Kosmidis and Passfield (2015) studied the training sessions of 14 well-trained long distance runners over a year. They gathered data generated by the running watches to calculate the effective speeds that contributed the most to their improved performances. A training distribution function was created using shape constrained regression analyses. The training distribution profile visualised the runners' over-all performances by expressing the amount of time that they spent at each achievable speed. The training distribution profile was then analysed in conjunction with performance outcomes. In this way they could identify the effective training speeds achieved in the field that made significant contributions to performance improvements (Kosmidis et al., 2015). By approaching the training analyses from the field, the researchers shifted training monitoring out of laboratory settings into the real world. The data covered a longer period of time that was most representative of real-world training and was not confined to the restricted laboratory spatial settings and availability.

Their work on the distribution curves for speed may be extended to the analysis of running form, that includes the geo-spatial gait metrics that are quantified by wearables and explained earlier in section 3.1. These metrics include cadence, ground contact time and vertical oscillation. These metrics may also be paired with changes in grade of the road, that is uphill, level or downhill running. A comprehensive view on an athletes physical running form during overground running may be presented to the runner, their coach and rehabilitation practitioners. This will enable the runner to identify their strengths and weaknesses during overground running, and how the environment influences their running style.

4. Risks of big data in sport: the reputability of research

There are some concerns relating to the ethical considerations of big data in sport. Three of them will be discussed here: data validity, data security and athlete autonomy. These factors may halt the progress of research that relies on tracking data from wearables, but may also become specialised research fields in their own right aimed at protecting users from harmful exposure of their private lives.

Data validity

Is the data accurate and does it correctly represent what it claims to represent? The reliability of sport tracking data is important as performance decisions are based on them. Incorrect readings will lead to over- or under-determination of performance capabilities and subsequently harmful decisions may be made. An athlete might push themselves physically to far or falsely assume fatigue due to some performance detriment (Karkazis et al., 2017). The veracity of the data is influenced by the commotion, or noise, that may accompany tracking data due to a faulty device, signal interruptions between the device and its interacting environment or the body of the user. Although a wearable presents the athlete with the opportunity to assess their physical performance, the user must still have the knowledge to distinguish between logical results and noise. Fanfang et al. (2013) tested some commonly used fitness tracking devices on their accuracy. For step count, the best performance of a device had an average 1.05% error rate with the poorest performing device having an average error rate of 27.28%. On measuring the distance the lowest error rate was 3.72% and the highest error rate 11.17% over 400 meters on a track (Fangfang et al., 2013). These statistics underscore the problem of veracity in the data. Error rates such as these will lead to inaccurate calculations of metrics such as cadence (stride frequency) and running speed or pace. An athlete may now believe they are under or over-performing and subsequently make unnecessary changes to their training or race tactics. Confusion and frustration may set in with the athlete and their coach, resulting in perhaps mistrusting the device and discontinue the use thereof.

Although algorithms designed to clean and present data should be objective, they still suffer from bias due to the perspectives and assumptions of the developer. The analysis of sport biometric data presents a unique challenge to algorithms: there is an overload of data that requires interpretation but an undersupply of historical, validated data to develop a valid algorithm. Somewhere data will have to be validated for algorithms to become reliable and true representatives of the actual data that is generated by a runner. The signal-to-noise ratio in biometric data remains low when compared to data collected in a structured experiment (Karkazis et al., 2017). Interpretation of the data must always be accompanied by some domain knowledge and not just taken as the full truth at face value. On the aggregate scale, such as the data used in the urban planning studies, data suffer from bias as it is limited to the population characteristics of those who are actually using the device. In the urban planning study for cycling volumes, the data from Strava is skewed towards male users at 87%. Nonetheless the extensively large scale on which the fitness application data can be collected outweighs the bias disadvantage (Hochmair et al., 2016).

Data security and protection

Personal data generated by fitness trackers are not fully and effectively protected, with as much half of data that requires protection are actually protected (Cortes et al., 2014). Hackers may gain access to sensitive personal information when a runner records their run session. For example, the starting and/or end location may indicate their home address. A runner's routine (i.e. when they are away from home, where they run, how long they are away etc.) may be revealed after some remote surveillance. Their health data such as heart rate and other biomechanics may also be disclosed by malware or cyber-attacks. The anonymised large datasets generated from wearables and extracted for aggregation purposes such as what was done in the urban planning study are not immune against data privacy risks. Despite anonymisation of personal data, it was found that 87% of the United States population may be identified by their zip code, gender and birthday (Sweeney, 2000). Wearables' tracking data amplify this risk: from GPS data it is easy to identify a runner's physical location at a certain time while their gender and birthday is captured on the on-line fitness profile.

Athlete autonomy

In the midst of the hype surrounding the advantages that wearables' big data can provide, a lingering question is surfacing: where is the line between data working for the athlete and data working against the athlete? Data analytics

should remain an adjunct tool in the athlete's quest for sporting excellence and career longevity, it should not become the driver of performance. Athletes risk loosing their autonomy and intuition when they completely rely on their biometric data to govern their performance efforts and race or game tactics. When an athlete is reduced to an entity basically consisting of only numbers and visualisations the enjoyment of the sport will be diminished, which is contradictory to the desired effect of sport participation in the first place for both the recreational and elite athlete. Sport fulfills the human body's inherent design to move at will, and not to be a marionette who plays by numbers.

The ethical dilemma: advances in research versus protection of human subjects
The World Medical Association Declaration of Helsinki ensures that where research involves human subjects, the study be subjected to ethical clearance by endorsed ethical committees. The ethical considerations as mentioned are a double-edged sword: though aimed to protect the athlete and preserve their privacy, it may also halt progress towards methods that can improve their physical health and access to infrastructure that support a healthy, active lifestyle. The ethical clearance process itself may be tedious and costly, prolonging the time it takes to execute experiments or other studies. This slows down the rate of progress in research fields with discoveries that could be advantageous to athletes, but at the same time prevent any suggested research where subjects may suffer unrepairable physical or psychological harm. Researchers in sport science and data science who make use of tracking data from wearables have an inherit responsibility to protect their subjects or participants from potential harm as well as preserving the integrity and quality of their research output. Advances in research in these fields must therefore be made in a responsible manner, even if it means a temporary decrease in momentum towards new discoveries or extension of the body of knowledge.

Trade-off decisions to advance or disprove of research suggestions are both complex and complicated, and can perhaps be compared to the complexities of traffic regulation: promote flow of transportation vessels, but also protect road users from accidents and harm by installing rules that limit continuous traffic flow. Not only must the ethical committee decide on the potential physical or psychological harm to participants caused by poorly executed research, but also take into consideration the protection of their personal data, online security, privacy and autonomy. Sport science research occurs in a multi-faceted environment, with big data extending this environment to include more variables and on a much larger scale. This union results in an intricate network of possibilities and/or problems that require advanced multi-disciplinary skills to solve, while keeping within ethical boundaries to protect the research subjects. Not all disciplines involved in data science have been exposed to the strict ethical considerations as have sport science or other medical research, so the collaboration efforts might face intrinsic challenges in how to work towards the common goal.

The risk of big data in sport therefor extends well beyond the damage that can be caused by cyber-attacks and data leaks. Data sets from wearables must be subjected to rigorous testing of integrity and preferably validated against known golden standards before adoption as tools to monitor physical performance. Researchers must present the true form of the data as it emerged as clear as possible, and disclose the veracity of the data together with the possible weaknesses of their models or other discoveries. Ethical committees are now faced with a new challenge, namely the trade-off between the acceleration of research made possible by big data and the protection of human participants in order to avoid costly mistakes made in the past.

5. Conclusion
The literature has shown that it is possible for not only sports disciplines to capitalise on new found knowledge hidden in the extensively large data sets generated by wearables, but also parallel research to advance physical activity within the population. The data from wearable tracking devices are representative of the real world and are not limited to a laboratory or clinical setting nor experimental time constraints. When used correctly this knowledge lends decision support to athletes, coaches, management teams and even urban planners that was not previously possible with the use of questionnaires, surveys and self-reported training loads. The discipline of data science

provides analysts with rigorous and scientific techniques to organise and transform the raw data into mathematical and statistical models. These models can be used for innovative approaches to training, tactics and urban infrastructure with the focus of improving physical activity by its inhabitants. However, the risks associated with big data must be considered in deciding which metrics are important and the extent of trust placed in the data. The wisdom gained from mining tracking data from wearables should provide athletes with a tool to make informed decisions, it should not govern the athlete. Researchers in the field must act responsibly with the data from wearables and respect the ethical considerations that accompany big data in sport science.

References

Balaban, Ö. & Tunçer, B. Visualizing and analyzing urban leisure runs by using sports tracking data. *City modelling tools*, vol.1**,** pp.533-535, 2017.

Cortes, R., Bonnaire, X., Marin, O. & Sens, P. Sport Trackers and Big Data: Studying user traces to identify opportunities and challenges. *Procedia Computer Science*, vol.52**,** pp.1004-1009, 2014.

CyclingSA. *Events*. South Africa: Cycling South Africa. Available: https://www.cyclingsa-events.co.za/app/ 5 May 2018. 2018.

Fangfang, G., Yu, L., Kankanhalli, M. & Brown, M. An evaluation of wearable activity monitoring devices. School of Computing, National University of Singapore, 2013

Hochmair, H. H., Bardin, E. & Ahmouda, A. Estimating bicycle trip volume for Miame-Dade county from Strava tracking data. The National Academy of Sciences, Engineering and Medicine: Transportation Research Board. 2016.

IDC. *Worldwide wearables market grows 7.3 percent in Q3 2017*. International Data Corporation. Available: https://www.idc.com/getdoc.jsp?containerId=prUS43260217 2017.

Iervolino, R., Bonavolontà, F. & Cavallari, A. A Wearable Device for Sport Performance Analysis and Monitoring. IEEE International Workshop on Measurements & Networking (M&N), Naples, 27 September 2017, 2017

Karkazis, K. & Fishman, J. R. Tracking U.S. Professional Athletes: The Ethics of Biometric Technologies. *The American Journal of Bioethics*, vol.17**,** pp.45-60, 2017.

Kaushik, A. *A Great Analyst's Best Friends: Skepticism and Wisdom*. Available: https://www.kaushik.net/avinash/great-analyst-skills-skepticism-wisdom/ 25 January 2018. 2016.

Kitchin, R. Big data and official statistics: Opportunities, challenges and risks. *Statistical Journal of the IAOS*, vol.31**,** pp.471-481, 2015.

Kosmidis, I. & Passfield, L. Linking the performance of endurance runners to training and physiological effects via multi-resolution elastic net. Cornel University Library, 2015

Mashooque, A. M., Soomro, S., Awais, K. J. & Muneer, K. Big Data Analytics and Its Applications. *Annals of Emerging Technologies in Computing*, vol.1**,** pp.45-54, 2017.

Morgulev, E., Ofer, H. A. & Lidor, R. Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, vol.1**,** pp.1-10, 2017.

Napier, C., Esculier, J. F. & Hunt, M. A. Gait re-training: out of the lab and into the streets with the benefit of wearables. *British Journal of Sports Medicine*. Published online, available: 10.1136/bjsports-2017-098637. 2017.

OMTOM. *Old Mutual Two Oceans Marathon History Since 1970*. Old Mutual Two Oceans Marathon. Available: http://www.twooceansmarathon.org.za/history 19 January 2018. 2018.

Passfield, L. & Hopker, J. G. A Mine of Information: Can Sports Analytics Provide Wisdom From Your Data? *International journal of sports physiology and performance*, vol.12**,** pp.1-7, 2016.

RunningUSA. *2016 State of the Sport: U.S. Road Race Trends*. Running USA. Available: http://www.runningusa.org/state-of-sport-us-trends-2015 24 January 2018. 2016.

Sands, W. A., Kavanaugh, A. A., Murray, S. R., McNeal, J. R. & Jemni, M. Modern Techniques and Technologies Applied to Training and Performance Monitoring. *International journal of sports physiology and performance,* vol.12**,** pp.63-72, 2017.

Statista. *Estimated size of the U.S. bicycle market from 2004 to 2015 (in billion U.S. dollars).* United States of America: Statista. Available: https://www.statista.com/statistics/255614/size-of-the-bicycle-market-in-the-united-states/ 5 May 2018. 2015.

Sweeney, L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University. 2000.

Wilkerson, G. B., Gupta, A., Allen, J. R., Keith, C. M. & Colston, M. A. Utilisation of Pratice Session Average Inertial Load to Quantify College Football Injury Risk. *Journal of strength and conditioning research,* vol.30**,** pp.2369-2374, 2016.

Biographies

Euodia Vermeulen is a Masters in Industrial Engineering student and assistant lecturer at the University of Pretoria, South Africa. She has earned BEng in Industrial in 2015 and BEng Honours in Industrial 2016. She received two awards as a final year student: SAIE Tuks best final year student for 2015 and the FNB most consistent academic achiever over the span of the degree. Earlier in life she has obtained a degree in Physiotherapy (BPhysT) at the University of Pretoria in 2010 and completed the community service year in Waterval Boven, SouthAfrica in 2011. Her research interests are in big data analytics and simulation, especially in the fields of medical, sports and health research and development.

Dr VSS Yadavalli is a Professor and Head of Department of Industrial & Systems Engineering, University of Pretoria. He received his Ph D from the Indian Institute of Technology, Madras in 1983. His research areas include, Applied Stochastic modelling towards Reliability Engineering, inventory management, Queuing theory and various estimation and optimization problems related to these areas. Professor Yadavalli has published over 130 research paper mainly in the areas of Reliability and inventory modelling in various international journals like, International Journal of Production Economics, International Journal of Production Research, International Journal of Systems Science, IEEE Transactions on Reliability, Stochastic Analysis & Applications, International Journal of Reliability, Quality and Safety Engineering, Applied Mathematics & Computation, Annals of Operations Research, Computers & Industrial Engineering etc. Prof Yadavalli is an NRF (South Africa) rated scientist and attracted several research projects. Prof Yadavalli is serving several editorial boards of various international journals. He supervised several M and D students locally and abroad. Professor Yadavalli was the past President of the Operations Research Society of South Africa. He is a `Fellow' of the South African Statistical Association. Professor Yadavalli received a `Distinguished Educator Award 'from Industrial Engineering and Operations Management Society in 2015. He also received a Silver medal for his research paper published in SAIMM in 2016.