Data Science

water e-journal
Online Journal of the Australian Water Association

ISSN 2206-1991
Volume 2 No 1 2017
https://doi.org/10.21139/wej.2017.008

LIFTING THE 'BIG DATA' VEIL

Creating Value through Applied Data Science

P Prevos

ABSTRACT

Big Data promises future benefits by using smart algorithms to improve the customer experience. Many organisations struggle leveraging the benefits of the data revolution. This paper summarises how the emerging field of data science can be used by water utilities to create value from information. The principles of data science are explained and illustrated using examples from water utilities. This paper closes with recommendations on how to implement data science projects to maximise value from data. These benefits are realised using existing investments in information technology infrastructure and existing competencies.

INTRODUCTION

The words "Big Data" have become synonymous with promises of virtually unbounded benefits. Big Data algorithms are attributed mystical capabilities in predicting the future. From improving the experience of customers, to optimising treatment processes, Big Data promises to profpoundly influence water utilities. There are successful examples of companies such as Facebook, Amazon and Google, where data science forms part of the fabric of the enterprise. But for most organisations, including water utilities, data science success has been limited to a few tests.

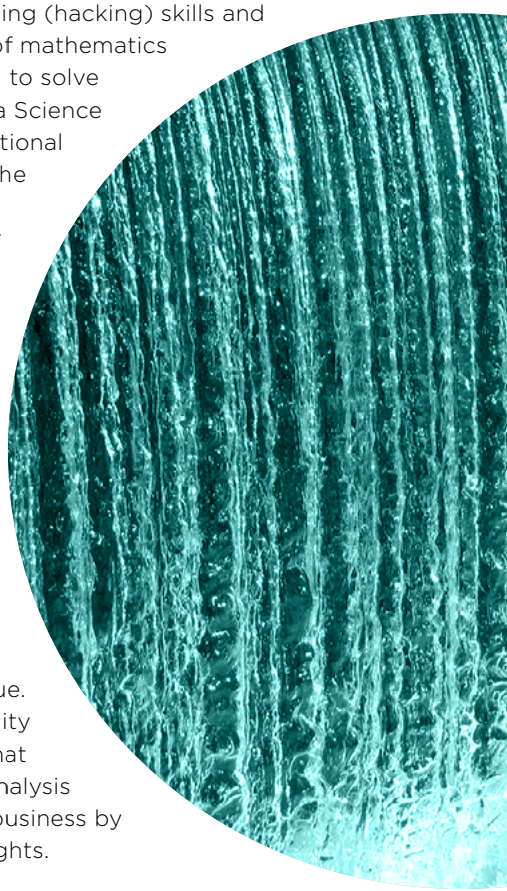
The envisaged benefits of Big Data have created a groundswell of interest in this topic within water utilities. This paper explains how water utilities can extract more value from existing data by using a strategic Data Science approach. This paper demonstrates how the benefits of Data Science can be realised by combining existing information technology infrastructure and competencies.

DATA SCIENCE

The term Data Science is more suitable to describe the process of creating value from data because the Big Data moniker is burdened with promise and hype. Data science is a systematic approach to analysing data. Although data analysis has always been the domain of engineers, new developments in information technology have turned this field into a specialised endeavour. Data Science is an emerging multidisciplinary field that exists on the confluence between knowledge of mathematics, coding skills and subject matter expertise. The difference between traditional analytical approaches and Data Science mainly exists in how data products are developed an integrated in everyday business.

The combination of skills required to undertake best practice data science are visualised in Conway's data science Venn diagram (Figure 1).

The combination of coding (hacking) skills and a good understanding of mathematics is a necessary condition to solve complex problems. Data Science moves beyond the traditional spreadsheets because the large volumes of data available for analysis far exceed the capacity of traditional tools. Data scientists use coding skills to develop databases and analytical software to manage the more complex tasks. These two skills need to be enhanced with contextual knowledge of the subject being analysed to be able to create value. Knowledge of water utility management ensures that the outcomes of data analysis add value through the business by creating actionable insights.



Business analysis undertaken by teams without expertise in water management can lead to outcomes that are not actionable due to a lack of context. Having said this, a fresh look on existing data can also open new areas of insight but given the technological complexity of water and wastewater services, subject matter expertise is required to make sense of data.

A data science team uses mathematical analysis to investigate a problem related to their area of expertise and uses computing skills to undertake and disseminate this analysis. The question arising from this introduction is how Data Science can add value to water utilities, beyond what is capable of achieving through standard methods?

Data Science for Water Utilities

The challenge to implementing data science in water utilities, which some call hydroinformatics, is how to transition the organisation from being data-rich but information-poor to making decisions based on insight backed by data.

Water utilities are ideal candidates to surf the digital revolution wave because they are traditionally data-rich organisations. Surveys conducted in 2015 with the Chief Information Officers from fifty large utilities in the United States indicate that only 10% of the available data is analysed to create value. The remaining 90% of the data, often referred to as 'Dark Data', provides a wealth of information that could be available if thoroughly analysed.

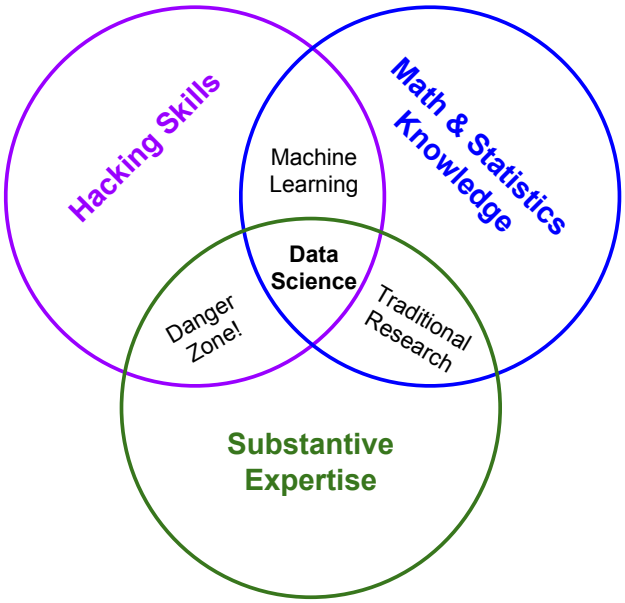


Figure 1. Conway's data science Venn-diagram (Conway, 2013).

To better understand the existence of Dark Data we need to separate data created for *ad hoc* operational purposes from data for *post hoc* analysis. In essence, Dark Data is a matter of context. Most of the data stored in operational systems, such as SCADA or the CRM, is used to assist the operational process.

As the interest for operational purposes wanes, this data becomes Dark Data. Most utilities own a large fleet of instrumentation that constantly measures a broad range of parameters through SCADA and other systems. This data is used to control core service delivery functions to manage the customer experience. There are many other opportunities to extract value from data after it has been used to manage operations. Data Scientists opportunistically use Dark Data for a purpose other than it was created for.

The purpose of integrating data analytics into an organization is to create value from data by providing *sound*, *useful* and *aesthetic* information, such as a report, an application, a dashboard, a plant automation algorithm and so on.

The *soundness* of the analysis requires the use of appropriate methods and the validation of results. The *usefulness* of data products is based on their ability to enhance the customer experience, reduce the environmental footprint of a water utility, improve the bottom line, or any other positive outcome.

Value is determined by whether the information provides actionable insights. Data science also needs to be *aesthetic* and follow the principles of best practice in data visualisation and reporting. The 'beautification' of the data ensures that the message is easily understood by those that consume the information and are thus more likely to make correct decisions much more rapidly.

THE DATA SCIENCE CONTINUUM

The art and craft of data science can be expressed in a continuum that shows business value as a function of the complexity and maturity of the analytics (Figure 2). These levels are hierarchal, which implies that to achieve the highest level of business value and maturity, all previous levels need to be progressed through.

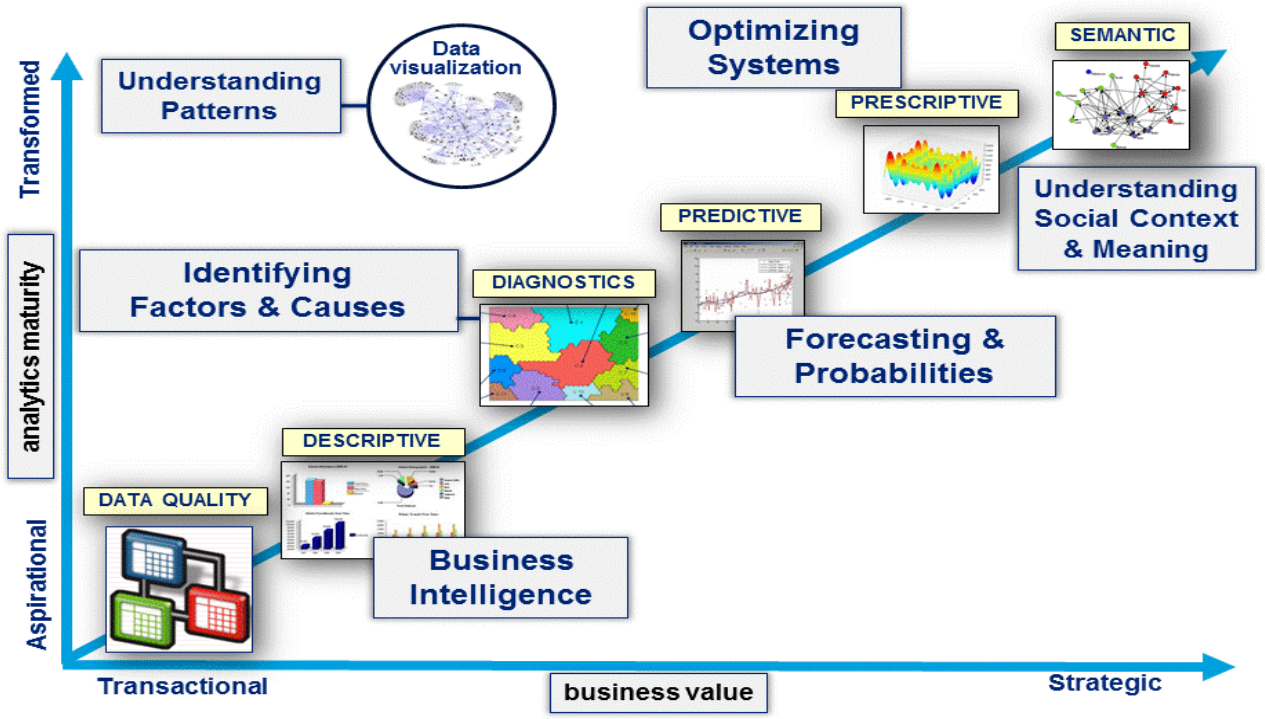


Figure 2. Data Science Continuum (Mongeau, 2014).

Data Quality

Data quality provides the underlying plumbing of the data science continuum. The majority of resources in any data science project are spent on cleaning and transforming data into a format that can be analysed. This work is not neccesarily a reflection of bad data management practices.

The main cause of this issue is that a most data is a by-product of operational processes. For example, a Customer Relationship Management system generates and stores data to facilitate the communication with customers, which is not necessarily in a format amenable to post-hoc analysis.

Data collected from SCADA Historians needs to be enhanced because the data is free of context. For example, filtered water turbidity data is generated 24 hours per day, but is only relevant when the filter is actually running. Two or more data sources need to be combined into one to provide meaningful information. At Coliban Water we have developed the Virtual Tag approach to extract and transform data form the SCADA Historian to make it suitable for Data Science projects (Prevos, 2016). The Virtual Tag engine currently contains data from bulk flow meters and critical components of the water treatment process and is used to detect non-revenue water and assess water treatment plant performance.

Descriptive Statistics

Most business reports consist of collections of descriptive statistics provided through tables of averages, maximums, minimums trends and other summaries. Descriptive statistics summarise existing data, but cannot generate new insights.

Descriptive statistics can be enhanced through visualisation techniques. The visualisation of data is an emerging field, where insights of graphic design and psychology merge to improve the way we consume information. Dashboards, infographics and other visualisation techniques help managers to quickly consume information created from complex data.

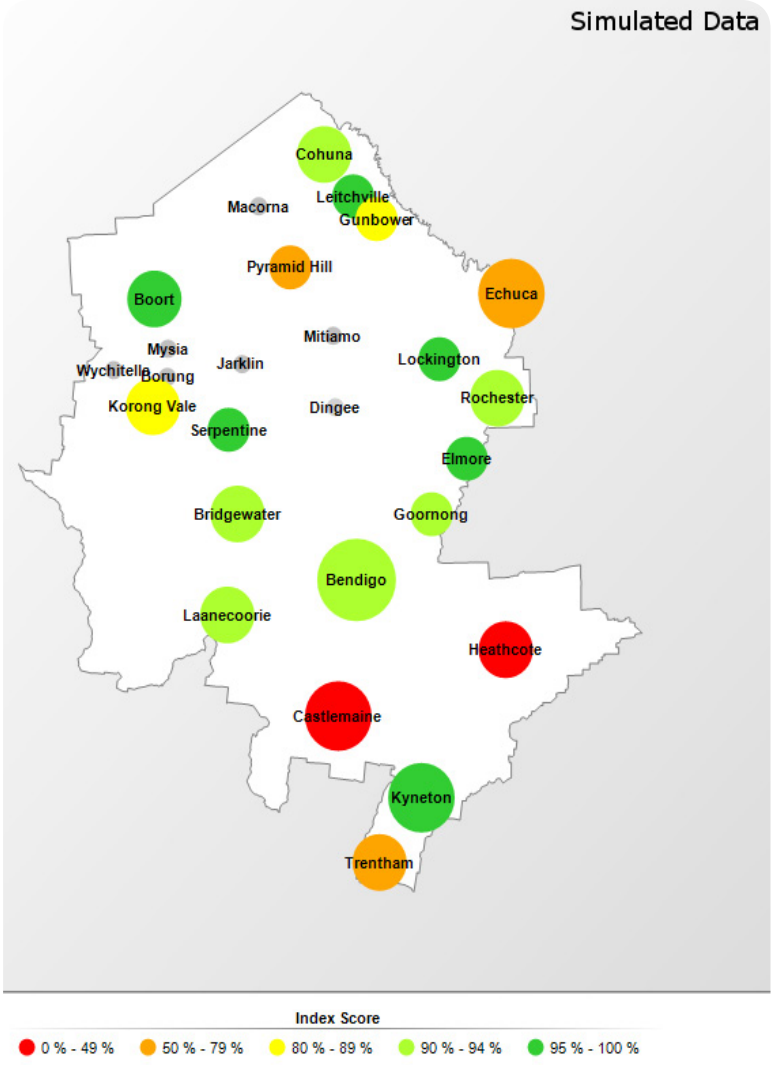


Figure 3. Water service index (Prevos, 2015).

At Coliban Water we have developed a dashboard to visualise water system performance to the Board. This index uses four different sources of information: CCP alarms, laboratory data, the register of regulatory breaches and customer complaints. This information is amalgamated into an index and visualised geographically. This report has moved the Board away from interpreting water quality data presented in numerical tables to asking meaningful questions suboptimal performing systems. The traffic-light map is clickable and all data sources and transformations can be interrogated in detail to perform a root-cause analysis.

Diagnostics

At the third level, analytics techniques are used to diagnose existing data and create new information. These methods are common in water utility management through the use of, for example, hydraulic network modelling or contact centre capacity planning. Analytics goes beyond traditional business intelligence, as it is aimed at creating new insights that are not present in the original data. Coliban Water has developed an automated methodology to implement the quantitative aspects of the microbial Health-Based Targets (HBT) manual published by WSAA. This system uses the previously mentioned Virtual Tags approach to add context to SCADA data and applies the decision rules in the HBT manual to assess treatment plant performance (Prevos and Sheehan, 2015).

Predictive

Most of the future value from Big Data will come from the third level of the data science continuum, which is associated with predictive analysis or machine learning. These algorithms are designed to detect patterns in data, including unstructured data such as customer interactions.

Predictive analytics can optimise asset replacement strategies, ensure sufficient staff are available in contact centres or optimise energy and chemical purchases. This is a rapidly emerging field that shows great promise for the water industry.

Prescriptive and Semantic

Prescriptive analysis uses the results from predictions to make decisions on behalf of people. This type of analysis regularly occurs in treatment plants, but the logic on which these controls are based is usually linear.

This is the level of Intelligent Water Networks, where predictive analysis is used to optimise operations.

Semantic analysis moves into the quantitative area of qualitative data and is used to analyse large volumes of text, social networks and other social data. When analysing complaints we are interested in extracting the voice of the customer from the data, which goes beyond simple statistics on complaint numbers.

CREATING VALUE FROM DATA

These philosophical considerations about data science need to be translated to business practice to create the promised value. The well-known Data-Information-Knowledge-Wisdom hierarchy categorises types of knowledge, but this triangle misses an important aspect. Underneath the data there is a reality that we seek to improve. Value is only created from data when the knowledge and wisdom is able to improve reality. Value is only created through actionable insights. Good Data Science is grounded in the physical and social reality which it aims to improve.

Water utilities are well-placed to embrace the new developments in Data Science because analysing data comes naturally to the engineers and scientists in our industry. Many of the competencies required to implement advanced analytics are already available to be utilised. One of the earliest examples of Data Science is in fact related to water supply. The famous cholera map drawn by John Snow in 1854 is one of the earliest examples of using data to improve public health.

Implementing data science in water utilities does not necessarily require large investments in software and external expertise. The 'R' and Python programming languages are Open Source tools with impressive capabilities in this area, used by many large corporations.

Most water utilities already have licenses for various Microsoft products, such as SQL Server Reporting Services, that can also be used to develop advanced Data Science products. Coliban Water is implementing a data science strategy based on the continuum in Figure 2, which has already delivered tangible results. We are currently developing an automated water balance for all our nineteen water systems and are paving the way to develop predictive models to help us optimise how networks are managed.

Coliban Water shares any intellectual property in this area freely with other water utilities to advance data science in this industry. The Health-Based Targets (HBT) software is currently shared with several other utilities under an Open Source license arrangement. The most effective way to obtain the benefits of Data Science within this industry is to pool intellectual resources to create better experiences for our customers.

THE AUTHOR

Peter Prevos is a civil engineer and social scientist with Coliban Water. He is responsible for creating value from data by providing sound, reliable and aesthetic information to assist the organisation in making better decisions.

REFERENCES

Conway, D. (2013). The Data Science Venn Diagram. drewconway.com/zia/2013/3/26/the-data-science-venn-diagram *(Accessed 2 November 2016).*

Lankow, J. et al. (2012). Infographics. The power of Visual Storytelling. John Wiley & Sons.

Mongeau, S.A. (2014). Emerging Trends in Data Analytics. sctr7.com/2014/07/09/twelve-emerging-trends-in-data-analytics-part-1-of-4 *(Accessed 2 November 2016).*

Prevos, P. (2015). Visualising water quality: A graphical index for drinking water system performance. OzWater, Adelaide.

Prevos, P. (2016). Health-Based Targets performance reporting: Virtual SCADA tags to facilitate data analysis. OzWater, Melbourne.

Prevos, P. and Sheehan, D. (2015). Health-Based Targets performance reporting. Water: Journal of the Australian Water Association (42)7, 62-64.

Water Services Association of Australia (WSAA) (2014): Drinking Water Source Assessment and Treatment Requirements. Manual for the Application of Health-Based Treatment Targets.

FIGURE SOURCES

Figure 1: Conway's data science Venn-diagram (Conway, 2013). Creative Commons, Attribution-NonCommercial.

Figure 2: Data Science Continuum (Mongeau, 2014, author contacted for permission).