IPENZ TRANSPORTATION GROUP CONFERENCE 2015

UNLOCKING DATA TO ACHIEVE SMARTER OUTCOMES

Authors

Michael Flatters, BEng (Hons), CEng (UK), MICE, MIPENZ
New Initiatives and Innovation Manager, MWH NZ Ltd
michael.j.flatters@mwhglobal.com

Dr Selwyn McCracken, Ph.D., MSc, BSc
Data Scientist, MWH NZ Ltd
Selwyn.b.mccracken@mwhglobal.com

Peter Hill, BSc
Data Analyst, MWH NZ Ltd
peter.n.hill@mwhglobal.com

November 2014

ABSTRACT

Effective transport management relies on making evidence based decisions. This means that decisions makers have to be able to access the right information or data, in the right way, at the right time and be able to interpret it in the right way.

Any disconnect between the information available, the data, and the decision making, can result in lost opportunities, reduced performance or erroneous decisions. Conversely, an overabundance of information that cannot be easily interpreted can result in a lack of clarity and also impede the decision making process.

This is where the application of data science can assist. Data science is a relatively new discipline that applies big data analysis principles and technologies to real business needs and outcomes.

Recognising the potential benefits of this discipline, MWH have established an Information Engineering Group focused on the four areas of Visualisation, Simulation, Prediction and Optimisation.

The paper will discuss the techniques used and the benefits these can bring to transportation – from public consultation through to asset management. It will briefly present a range of demonstrable examples where these techniques have been applied to provide smart solutions to our clients.

INTRODUCTION

The efficient transportation of people and goods is an essential component of the modern economy. However, transport also poses many difficult challenges because of its central importance, scale and often global impact. Moreover, the inter-connecting nature of transport means that many issues overlap, frequently in conflict with one another, and can simultaneously concern several geographically dispersed communities and stakeholders. Specific issues can include:

- Energy concerns about efficiency, fuel type, cost volatility and the security of supply and distribution
- Pollution effects on the environment and health
- Sustainability of infrastructure development, management and maintenance
- Maximising trade, productivity and economic growth
- Improvement of social equity, community connectedness and cohesion
- Identification of the best project from a range of options
- Effective and meaningful public consultation
- Ensuring competition and effective market regulation
- Minimising safety risk
- Transport network resilience under adverse conditions
- Planning for population growth and demand
- Logistics and supply chain management
- Reduction of traffic congestion
- Integration of new systems and technology

Our fundamental premise is that complex problems like these are most effectively attacked by the collection, integration, analysis and interpretation of relevant data. We also recognise that it is much easier to collect data than to extract coherent knowledge from it, especially as new technologies generate ever larger volumes of data. Nonetheless, the size of reward can be enormous. For example, a 2013 report by the OCED estimated that by 2020, the analysis of data acquired through the tracking of mobile devices could deliver USD $500 billion of value worldwide in the form of fuel and travel time savings, or a reduction of 380 megatonnes of CO2 emissions. [Ref 1]

As the scale of data balloons however, increasingly specialised skills are required, namely that of data science. This paper will explain why data science has become increasingly valuable. It will touch on big data and what is meant by data science. Using international and local examples, it will show how it has been applied to the transportation sector overseas and here in New Zealand.

DATA, DATA EVERYWHERE

Data can be defined as raw unorganised facts. Data is everywhere, and the volume, variety, and velocity of it continues to grow [ref 2]. Recent reports cite that 90% of the world's data has been created in the last two years [ref 3]. This amounts to several trillion bytes of data per day [ref 4].

However, only when data is processed, interpreted, organised, analysed and presented does it become meaningful. When this occurs, data is transformed into information and this information can then assist with decision making.

Examples of raw data within the transportation discipline range from traffic counts to the physical dimensions of a road. However, there are many other sources of untapped data that are not currently being utilised but may ultimately assist with decision making.

BIG DATA

Figure 1 Unlocking Data

The term 'big data' refers to data sets of extreme size, diversity and complexity. Every day the size of these data sets increases through interactive transactions such as credit card sales, mobile phone use or communication through social media.

So what exactly defines 'Big' data? The McKinsey Institute deliberately and sensibly avoid a definition that quantifies a threshold of big data because of the rapid evolution of technology, both in terms of data generation and processing power. Instead, they define 'big data' as "datasets whose size is beyond the ability of typical database software to capture, store and analyse" [ref 5]. They nonetheless state that big data for many sectors will be in excess of a few dozen terabytes up to multiple petabytes in size.

In terms of a practical definition, we consider 'big data' to arise when more than one standard workstation computer is required to store and process data.

There are many large data sets relevant to the transportation sector - some of which are already available, others that are not currently being collected but would be useful. Some of these include:

- Traffic Movements (source, route and destination)
- Traffic Interactions (travel times and delays)
- Economic Use of Networks (transport vs others uses)
- Economic Cost of Networks
- Environmental Factors (emissions)
- Resilience / Risk Issues
- Asset Data (what, where and condition)
- Geographic Issues (weather, terrain – lidar)
- Social Factors (ease of use)

This list is certainly not exhaustive but provides an idea of the size and complexity of some of the data that could be of value. Some datasets may in fact not be considered 'big', but in combination with all others they most certainly would qualify.

Hidden within these datasets are many sources of untapped information or relationships that are not currently visible without combining and analysing this data. With the right skills and techniques, innovative insights and information can be extracted from these big data sources to further assist with making decisions.

The McKinsey Institute reports five ways that big data can add value: [ref 5]

1. It can unlock significant value by making information transparent and usable at much higher frequency.
2. As organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore exposes variability and boosts performance. Leading companies are using data collection and analysis to conduct controlled experiments to make better management decisions; others are using data for basic low-frequency forecasting to high-frequency nowcasting to adjust their business levers just in time.
3. It allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services.
4. Sophisticated analytics can substantially improve decision-making

5. It can be used to improve the development of the next generation of products and services. For instance, manufacturers are using data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance (preventive measures that take place before a failure occurs or is even noticed).

Many of these are applicable, in some form, to the transportation sector and could be specifically applied to:

- Controlling Traffic
- Transport Planning and Modelling
- Route Planning
- Congestion Management
- Intelligent Transport Systems
- Route Planning and Logistics
- Revenue Management / Funding Sources
- Asset Management

DATA SCIENCE

Unlocking insights hidden within data, big or small, is the vocation of the relatively new discipline of data science.

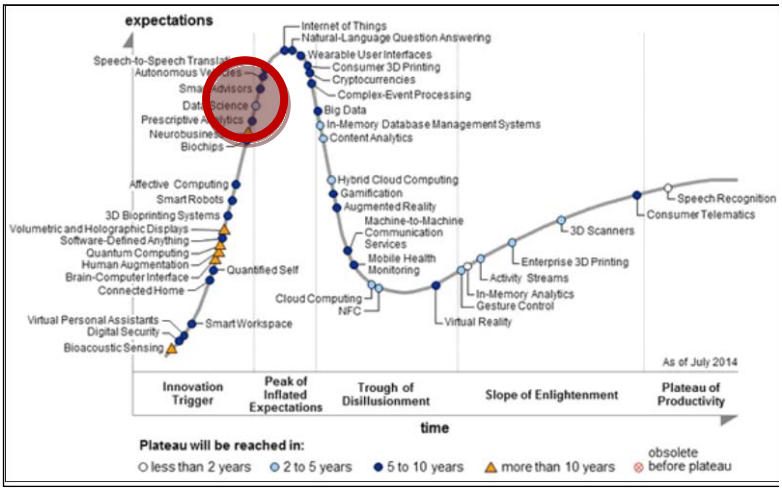

Figure 2 Gartner Hype Cycle

Data science was a new entry on the 2014 Gartner Hype Cycle [ref 6], an annual report that evaluates the market promotion and perception of value for over 2,000 technologies, services and trends (refer Figure 2). Technologies are shown relative to their expectations and estimated timeframe until they develop into a truly useful productive state.

Forbes [ref 7] described data science as 'more a discipline for dealing with big data then a specific technology or set of technologies' so it's inclusion on the hype cycle is interesting in itself.

So, what is data science and what skills does a data scientist bring?

A data scientist combines a broad range of skills [ref 8]. These include

- Collaboration and communication skills to understand stakeholder and business issues
- Analytical and decision modelling skills for discovering relationships within data and patterns
- Data management skills to build relevant dataset used to undertake the analysis.

Figure 3 shows the mix of maths and statistics, hacking skills and subject matter expertise required by data scientists. Expanding on this, the following list outlines the range of skills required by data scientists: Ref 9

- Analytical skill-set
- Mathematics / statistics (including experimental design)
- Domain knowledge (i.e. Industry specific processes where analytic are applied)
- Technology / data
- Communication skills (story-telling)
- Curiosity (willingness to challenge the status quo)
- Collaboration
- Commercial acumen/ Strategic
- Customer-centric
- Problem-solving skills
- Proactive

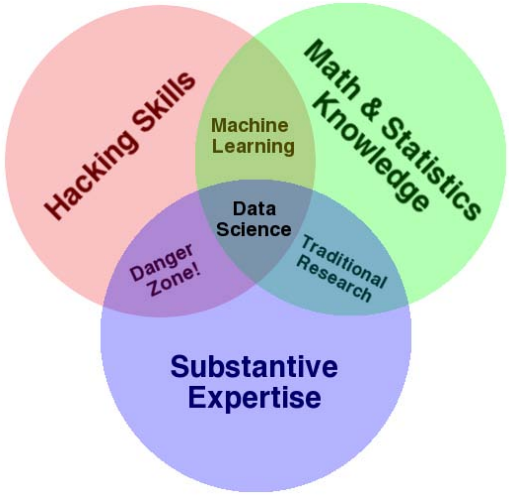

Figure 3 Required skills of a Data Scientist

From our perspective, the key elements of data science span the following techniques:

- Visualisation – The presentation of data to communicate information clearly and effectively
- Simulation – Representation of real world processes using quantitative models to examine a systems behaviour
- Prediction – Forecasts about the likelihood of events under any given scenario
- Optimisation – The selection of the best course of action from a range of alternatives, after consideration of various criteria.

These elements are clearly wide-ranging and not part of the traditional transportation skill set.

OPEN DATA

With the increasing volume, velocity and variety of data being produced, the transportation sector needs to consider how it can maximise the benefits from this skill set and what sources of data are required. So, how can this be achieved?

To answer the above question a couple more questions need to be answered.

- Does the transportation sector deal with big interconnected data sets?
- Are there further insights or information that could be gained from interpreting, organising, analysing and presenting data in a better way?
- Is all the data readily available to allow any analysis to be completed?

The answer to the first two questions is obviously yes, but the third probably best answered as perhaps or partially.

At the present time much of the asset data is ‘locked’ in proprietary asset management systems that restrict access to much of the data. Other useful data may be locked in to client systems or owned by others. At the present time, even basic transportation data in New Zealand is locked in this way. This makes it difficult for those outside of the industry to access this data to gain any

insights from the information they could contain.

Figure 4 New Zealand Government's ICT site

However, there is a direction from most government's internationally, and some private companies to move towards releasing public data for public use. This philosophy of 'open data' and the potential benefits that it brings have been recognised by the New Zealand Government and forms an important part of their Information and Communication Technologies (ICT) Strategy [ref 10]. This 'open data' can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike

The Minister for Land Information, Michael Woodhouse, reinforced this in the 2014 Report on Agency Adoption of the Declaration on Open Government Data. On its release he stated "government agencies are increasingly releasing public non-personal data in open formats for reuse," and that "In turn, third parties are using that public data in increasingly innovative ways – creating a raft of new products, tools and services for use by industry and the wider public

The recognised outcome of releasing this data is innovation.

With the combination of big open data sources and data science some real insights can be gained from the raw data.

EXAMPLE APPLICATIONS TO TRANSPORT

There are many practical examples of the application of data science across many industries and the benefits that it can bring. A couple of these relevant to transportation are detailed below.

Example 1 – AllAboard - a system for exploring urban mobility and optimizing public transport using cellphone data

An interesting example is the 'Data for Development Senegal' [ref 11]. This is an innovation challenge utilising open source big data for the purposes of societal development. Mobile phone companies Orange and Sonatel have released anonymous statistical databases and samples extracted from the mobile network signals in Senegal. This raw data has been made available for the 'D4D Challenge' that has been set up to provide socio-economic benefits in Senegal.

The challenge is focused in the areas of

- health
- agriculture
- transport/urban planning
- energy
- national statistics

A previous challenge winner was IBM's AllAboard project [ref 11]. This project determined travel patterns from individual call location data to propose new bus routes. An optimisation model was used to that effectively redesigned the public transport network. The result of this was improved user numbers, improved travel times and decreased waiting times.

Example 2 – DataKind – Out on a Limb [ref 12]

New York City's Parks Department is responsible for about 600,000 trees. As these trees get old they can present risks to properties, traffic and pedestrians. The department wanted the following questions answered

Figure 5 Central Park Trees – New York

- Does preventative tree care in one year reduce the number of hazardous tree conditions in the following year?
- Can we predict where trees will be most vulnerable to a storm and plan accordingly to minimize post-storm work?

The department already had a lot of data from a multitude of sources, across different time periods, but not all of it was organically compatible. Through the application of data science the following conclusions were determined

- Prudent pruning shrinks emergency clean-ups by 22 percent.
- Trees pruned every five years, as opposed to 10 or 15, pose less risk.

An interactive map was also produced that quantified the potential storm risk of the trees for future planning purposes.

The above examples applied the four components of data science to gain insights from the available data. AllAboard utilised simulation and optimisation, whilst DataKind applied prediction and visualisation techniques.

APPLICATIONS TO NEW ZEALAND TRANSPORT SECTOR

The Information Engineering Group of MWH is applying data science techniques to current client issues in the transport sector. Below are examples for each of the key elements of data science that we think are essential, namely visualisation, prediction, simulation and optimisation.

Visualisation

Problem – Unlocking information contained in large related crash and road asset databases.

Solution – An interactive reporting tool that show information spatially and allows results to be filtered against selected criteria.

Benefit – Improves efficiency as access to information is readily available and interpreted.

Details – New Zealand's Crash Analysis System (CAS) is a specialised system maintained by the Ministry of Transport. This system requires extensive training to fully utilise and extract data from within it. Consequently, there are few people that are able to use CAS to its full potential. Furthermore, 3rd party stakeholders such as local authorities are unable to easily examine crash information, trends and patterns that pertain to their area without that support. Additionally, the CAS data does not include information about the local road conditions that may contribute to each

crash.

For this reason, MWH has developed a revised web-based interactive mapping tool to more easily query CAS data. This allows users to interactively drill down and explore crashes in an intuitive and widely accessible manner. It also merges road surface condition data from RAMM where available.

Figure 6 MWH's interactive crash mapping site

Simulation

Problem – Cost risk uncertainty for a complex project

Solution –Overall and item-specific risk profiles obtained via Monte-Carlo simulation

Benefit – Improved confidence with decision making and elimination of 'gut feel' risk assessment.

Details - Every construction project contains several elements of risk, such as cost over-runs, missed deadlines and failure of structural components. However, with large, complex projects it is often a challenge to understand the overall project risk profile and how each project component contributes to that overall risk.

Monte-Carlo simulation is a powerful statistical technique that calculates a project's overall risk profile by combining the uncertainty profiles of each project component. For this reason, NZTA require risk analysis and contingency assessments to be completed for many new road projects using Monte Carlo techniques, as specified in their Cost estimation (SM014) and Risk Management (Z44) manuals. However, undertaking these analyses typically requires the use of expensive specialised software such as @Risk, which makes it difficult to collaborate and compare scenarios with those who do not have a licence for the same software. To address this problem, MWH has developed a web-based Monte-Carlo tool which allows for widespread sharing of models. This tool has been successfully used within MWH to assess exposure to cost-risk on projects, for example as part of a tender for a Network Outcomes Contract with NZTA. This tool will be released publically later in 2015.

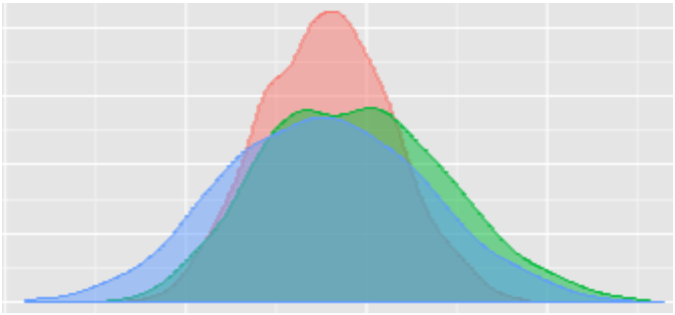

Figure 7 Monte-Carlo simulation profiles

Prediction

Problem – Asset replacement forecasts - lives of surfacing seals on the state highway network.

Solution – Statistical analysis of large data set to produce of survival curves for individual surfacing types.

Benefit – Improved decision making as confidence levels relating to the data are better understood.

Details – Most assets such as roads eventually require replacement or repair. Consequently many organisations want to forecast their future replacement and repair costs based on historic failure rates. Additionally, by understanding the risk factors associated with past failures, intervention activities can be deployed to extend the lives of surviving assets, or design improvements can be made before new assets are deployed.

For one recent MWH project, the survival profiles of different road surface materials (e.g. Asphalt, Chip seals, OGPA) were examined across the 11,000km of state highway network of New Zealand. For this network, the longevity of each material was established after consideration of factors such as traffic volume, proportion of heavy vehicles and region. Of note, by considering data from all road layers, that is surviving and replaced, substantially more accurate survival curves were developed than the client's existing method which completely ignored the top layer.

With accurate survival curves to hand, it was straightforward to estimate how many times a particular surface material would need to be replaced over 5 to 15 year planning horizon. This in turn allowed the cost-effectiveness of each material to be assessed against the others, given the differences in cost and anticipated replacement rates.

For this project, the RAMM database was the sole source of data. The degree of unexplained variance however, made it apparent that future work will need to incorporate other sources of data such as sub-surface conditions, both soil and geology, local area precipitation and surface material quarry sources.

Optimisation

Problem – Planning inspections of a large number of assets on a dispersed network.

Solution – Production of an optimisation model that showed the most efficient way to visit all the assets in the shortest time. Constraints were built into the model that included travel time and inspections duration.

Benefit –A potential overall cost saving of 27% when compared to the actual inspections undertaken.

Details - The routing of vehicle fleets is a common transportation logistics problem. The essence of the Vehicle Routing Problem (VRP) is to find the fastest set of routes that a fleet must travel to complete a set of tasks under some operational constraints. Typical VRP constraints include:

- Delivery time windows
- Vehicle capacity
- Customer demand
- Service times

Numerous applications of VRP techniques exist, such as:

- Travelling salesperson/inspector/courier route optimisation
- Road maintenance cost estimation (for road contractors)
- Pick-up and delivery problems (e.g. school bus route plans)
- Fleet size and mix planning
- Depot location evaluation
- Staff/vehicle rosters (e.g. scheduling driver rests & vehicle repairs)
- Stock inventory management and resupply
- Periodic monitoring routes (e.g. Police patrols)

The effective application of VRP techniques can lead to substantial efficiency savings in fleet running costs. For example, a 27% reduction in travel time was identified for an MWH project to inspect approximately 200 bridges across Fiji. This savings was found while still respecting the operational constraints of the project, such as length of working day, maximum travel speeds and current depot locations.

CONCLUSION

The amount of data available to help make effective decisions is very large and increasing. The examples presented have shown that the application of data science to large raw data sets can yield information that is meaningful and useful to stakeholders at all levels of the transport sector, from daily operations to long-term strategic planning.

It is also clear however that there are substantial barriers, beyond just technical skill ability, to obtaining and integrating traditional and novel sources of transport data. Many core datasets are not readily available beyond the transportation community, or are privately owned, meaning that useful data is underutilised or is inaccessibility to interested parties. For this reason, the concept of open data has been recognised as an important innovation driver and forms part of the government's ICT strategy, and should be embraced by the transport industry

Irrespective of these barriers, we firmly believe that the analytical techniques of data science can add significant value to the transport sector.

REFERENCES

1. *Internet*
   Exploring Data Driven Innovation as a new source of growth – OECD Report
   DSTI/ICCP(2012)9/Final – June 2013Internet
   http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP(2012)9/FINAL&docLanguage=En

2. *Internet*
   3-D Data Management: Controlling Data Volume, Velocity and Variety
   http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

3. *Internet*
   Big Data, for better or worse, Science Daily
   http://www.sciencedaily.com/releases/2013/05/130522085217.htm

4. *Internet*
   Data for Development - Orange
   http://www.orange.com/en/about/Group/our-features/2013/D4D/Data-for-Development

5. *Internet*
   Big data: The next frontier for innovation, competition, and productivity - McKinsey Global Institute
   http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

6. *Internet*
   Gartner Hype Cycle
   http://www.gartner.com/technology/research/hype-cycles/

7. *Internet*
   It's Official: The Internet Of Things Takes Over Big Data As The Most Hyped Technology
   http://www.forbes.com/sites/gilpress/2014/08/18/its-official-the-internet-of-things-takes-over-big-data-as-the-most-hyped-technology/

8. *Internet*
   IT Glossary, Data Scientist, Gartner
   http://www.gartner.com/it-glossary/data-scientist

9. *Internet*
   Data Science Central – Michael A Sanders
   http://www.datasciencecentral.com/profiles/blogs/data-scientist-core-skills

10. *Internet*
    Transforming Government ICT
    http://ict.govt.nz/strategy/

11. *Internet*
    D4D – Data for Development
    http://www.d4d.orange.com/en/presentation

12. *Internet*
    AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data
    http://www.orange.com/en/about/Group/our-features/2013/D4D/Folder/best-development

13. *Internet*
    Out on a Limb – Datakind
    http://www.datakind.org/projects/out-on-a-limbfor-data/