# The Alan Turing Institute Programme on Data Centric Engineering's Response to the National Infrastructure Commission's Second Call for Evidence

## 1. Introduction

**Background**

The Alan Turing Institute is the national institute for data science, with a mission to make great leaps in data science research to change the world for the better. The programme on data-centric engineering (DCE) will develop critical data analytic capabilities to address the challenges in improving the performance and resilience in engineering systems and national interdependent infrastructure nexus. The evidence presented in this document will be based on the Turing Program for DCE will focus on 3 grand challenge areas of:

1. **Resilience:** Resilient and Robust Infrastructures,
2. **Monitoring:** Monitoring Safety of Complex Engineering Systems,
3. **Design:** Data Driven Engineering Design under Uncertainty.

## 2. Better Asset Management

**Summary and Response to Questions:** The application of new data analytic methods in critical infrastructure (CI) asset management can significantly improve the efficiency, consumer experience, and reduce operating costs. The immediate technology priorities for efficient algorithm scalability and integration into asset management methods and tools include: sparse data combining methods, high-dimensional statistical inference models, and to automated data wrangling. Potential barriers to rollout include the: a) differential attitude and readiness to adopt new technology due to geographic or political segregation in certain CI operators (e.g. water management), b) complexity and uncertainty in integrating new methods into large-scale existing practices, and c) data privacy restrictions. The latter two cases are especially a concern for large CI operators. To overcome these barriers, we hope the government can encourage the uptake of new data-driven solutions through raising awareness of projects and the targeted investment in joint academic-industry research grants, with a focus in breaking down CI silos and unlocking data access restrictions. Certainly, the development of a national digital twin can help to: (1) overcome the barriers by bridging geographic and sectorial divides through linking interdependencies via a common model, (2) provide a framework for determining sensor locations, and (3) serve as a technology demonstrator for new tools. As such, the digital twin must connect disparate CI sectors and be open to the demonstration of new data analytic tools.

**National Importance:** UK infrastructure is ageing, and requires an ever-increasing amount of investment in maintenance and upgrade to maintain existing performance levels. In addition, infrastructure assets are characterized by long life and complex deterioration modes; knowledge about the way these assets deteriorate over time and how the deterioration affects the risks and asset performance is patchy. In summary, today's infrastructure is faced with familiar and seemingly insurmountable problems – too little money, too many assets and increasing complexity.

We now give **examples of projects** underway in using data analytics to improve engineering assets in **critical infrastructure sectors**:

- **Flood Risk in London Underground:** Transport infrastructure assets are at risk from changing environmental conditions, in part contributed to by climate changes observed in recent decades. In the U.K., increased precipitation amounts are leading to rising groundwater levels, presenting transport operators with performance issues associated with flooding of rail tracks and ballast. A NERC funded feasibility study (NE/M007987/1) derived a groundwater rise vulnerability model for the cut-and-cover tube tunnels bearing on terrace gravel deposits in London Underground (LU). This sets out the mechanics of the seepage problem and through a deterministic approach has identified upper and lower risk boundaries considering a groundwater level fluctuation range [1]. The autonomous systems developed will enable cost-efficient continuous monitoring strategies to be put in place, and are especially valuable at network scale where interconnected assets require simultaneous inspection to fully understand the risks. Once analyzed, these will provide a probabilistic understanding of the risks posed to asset performance through robust hazard-structure interaction modeling. This data-driven approach to asset management will provide LU with high-resolution performance statistics that reflect the risk model in place. This allows optimisation of drainage, ballast and track maintenance through predictive strategies not currently achievable with traditional inspection regimes. This in turn empowers LU into shaping and optimizing the resilience of these assets for the future.
- **Self-Organisation in 4G Networks:** The ICT industry is a leading producer of digital data and has for decades used its own data to automate asset management. In recent years, there is a growing recognition to combine ICT data with new forms of social media and mobile data to create stronger user-centric understanding of consumer demand and consumer experience [2]. To do so, joint academic and industrial initiatives are underway (EU H2020 project 778305, InnovateUK project 010734) to transfer state-of-the-art heterogeneous big data analytics and machine learning tools into applied ICT automation algorithms in critical industries such as 4G/5G mobile networks. The analytical techniques involve high-dimensional statistical models using Gaussian Processes and Deep Learning to forecast heterogeneous data demand, as well as stochastic multi-armed bandit algorithms with performance guarantees to drive a range of automated asset management across time scales (millisecond resource assignment to daily asset adjustments). These form the important building blocks to virtualise asset management and reduce OPEX in current and future networks (5G).
- **Railway Infrastructures:** Asset management in the rail industry is critical. For example, in the financial year 2009/10, whole-industry costs totaled £12.7bn. Of this, over half was spent on maintenance, renewals and enhancements. National Rail own 30,000 railway bridges and considerations are underway to instrument the bridges, yielding a data storage and analysis bottleneck. What is required are on-the-fly procedures that can be employed without storing all data. Currently, a collaboration is formed between the DCE at the Alan Turing Institute and the Cambridge Centre for Smart Infrastructure and Construction at Cambridge University, and Imperial College to develop *intelligent digital twins* for two railway bridges (in collaboration with Laing O'Rourke Plc. as part of the Staffordshire Alliance Improvements

Programme - SAIP). The instrumentation of bridges has changed the hands-on assessment of a bridges behavior to include a statistical data analysis. A statistical model will give an understanding of the stochastic nature of bridges and lead to an efficient monitoring system for predictive maintenance. The combined use of statistical analyses, big data, physical modelling and numerical modelling constitutes the main features of the digital twin. The real world SAIP self-sensing bridges are serving as the training ground for validating the intelligent digital twins. The approach, if employed over a system of assets, enables asset managers to (1) to develop a novel whole-system asset monitoring and maintenance capability and (2) get appropriate and accessible asset information that enables timely and cost-effective decision-making at different times of the assets' lifetimes.

## 3. Smart Traffic Management

**Summary and Response to Questions:** Smart traffic management (STM) has already started. There are 3 decisive challenges STM systems may have to tackle.

1. STMs will need to collect and process real time, high quality data.
2. increased demand for individual transportation will need to be offset by improvements in traffic flow density.
3. It will be necessary to improve the efficiency of movement via increased ride sharing and interchangeability between modes (i.e., improved coordination).

Access to relevant data is a key contemporary challenge for urban policymakers as they deal with ever growing demand on public infrastructures and considerable financial constraints. As part of its work on data centric engineering, the Oxford Internet Institute (OII) in conjunction at the Alan Turing Inst. have been conducting research into the deployment of open and social media data for facilitating smart urban management. The key aim of this strand of work is to find ways of enabling an internet of things style awareness of the surrounding urban environment without the up-front costs and difficulty of installing large sensor grids (which are out of reach of all but large urban megacities): instead, we are exploring ways of repurposing existing data created by third parties and government itself. This creates what we have described as a "lightweight" smart city [3].

We now give **examples of projects** underway in using data analytics to improve traffic management in different sectors:

- **Data Bias from Public and Social Data:** The research frontier in this area concerns plugging missing traffic data with repurposed data and accounting for the biases. Open data has the potential to change the way we collect and process transport data. One of the most promising projects is the open data platform created in the city of Manila. Easy Taxi, Grab and Le.Taxi – three ridesharing companies – partnering with the World Bank are sharing their driver's GPS streams to the public using an open data license. This Open Transport Partnership makes it possible for transport agencies to make real time evidence-based decisions at relatively low cost. Examples of recent work by the OII including the use of OpenStreetMap data for understanding the spatial availability of alcohol [4], and Twitter data for understanding local high resolution commuting patterns [5]. In so doing, we highlight two key findings. First, there are biases in the demographic makeup of the groups which contribute to open and social media platforms. Second, we have found that these biases are not so severe that they impede the extraction of reliable proxies, which were found in the case of both alcohol availability and local commuting patterns.

- **Demand Mitigation using Autonomous Vehicles:** Autonomous cars may lead to higher congestion due to demand effects from the forecasted substantial drop in the monetary costs of travelling by car. To mitigate demand externalities, traffic efficiency gains will have to be maximized. Depending on the level of automation, substantial gains could come from smart lights. Recent works shows that a simple light system with human drivers could increase traffic flow efficiency up to 200%. Autonomous cars make it possible to move from the traffic flow based system to a vehicle level system. This could substantially increase capacity and significantly reduce delays at intersections. Another way to respond to increased demand will be through improved coordination, and specifically, via ride-sharing and better interchangeability between modes. The technological innovation of smart phones and the decreasing cost of computing made it possible to efficiently share rides. Researchers at MIT have created a model that predicts the potential for ride sharing in any city. This potential, measured by the compatibility of individual mobility patterns in space and time, is shown to be substantial with important implications for demand management.

## 4. Big Data

**Summary and Response to Questions:** The effective use of big data requires greater standards to make the data accessible and usable. Currently data from numerous sources will be in various states of readiness, and combining datasets and getting value from them in an arduous task. This would be made easier by having defined and widely accepted standards for data structures, data labelling, data cleanliness and data-sharing methods. The Alan Turing Institute is working with industry and public bodies on the development of standards for data science and on defining Data Readiness Levels to better methodology is how big datasets are managed. Also vital is the widespread acceptance of appropriate data security procedures. Many company are failing to protect vital infrastructure data, through reliance on lax procedures or outdated hardware and software. Solutions to these problems exist, but are not being adopted enough.

**Open Data:** The benefits to having open data in the modern age are unprecedented, especially where they impact public services. Open data and accessible APIs can lead to greater public awareness and engagement with infrastructure, new services, greater safety and gains in efficiency. It also opens the sector to greater innovation from data science firms, especially the UK's wealth of start-ups and SMEs in this space. However, the sector as a whole is unwilling to share data openly, and even private data-sharing agreements (B2B, collaborations with academia, etc.) can be difficult to arrange. The unwillingness to openly share data is largely a cultural issue stemming from conservatism in many parts of the infrastructure business sector. There is a fear of the implications of sharing data openly, particularly around legal ramifications, security considerations and loss of IP. Many of these fears stem from a lack of knowledge and experience in operating with open data. Possible solutions to address these issues include:

- Government guidance on data-sharing methods, including standards for ensuring security of data structures and advice on adhering to legal restrictions around data protection and other data-related legislation.
- Flagship schemes or pilot projects to show the value and potential of data-sharing initiatives. This could build on existing schemes, such as the use of APIs by TFL for tube and bus services which has led to a range of improvements for customers travelling by public transport.
- Financial incentives for firms which engage in open data sharing
- Regulatory incentives which can nudge companies towards sharing data

**Digital Twin:** A national digital twin is important to provide modeling and forecasting to an ageing UK infrastructures. It should provide a platform for using data and data science to validate and reinforce existing mathematical models of complex engineering systems and assist in the development of new models. Certainly, the development of it can help to: (1) overcome the barriers by bridging geographic and sectorial divides through linking interdependencies via a common model, (2) provide a framework for determining sensor locations, (3) identify abnormal behaviour using machine learning techniques that do not expose it to adversarial attacks [6], and (4) serve as a technology demonstrator for new tools. As such, the digital twin must connect disparate CI sectors and be open to the demonstration of new data analytic tools. Only by bridging data (collection, analysis) and engineering knowledge, working with engineers and knowledge stakeholders, can a national digital twin help to manage both the data and the infrastructure in an efficient and reliable way.

**References**

[1] Stephenson, V.; D'Ayala, D. 2017. Assessing the Vulnerability of Historic Rail Tunnel Linings to Groundwater Rise, Quarterly J. of Engineering Geology and Hydrogeology

[2] Fan, C. et al.; (2017) Learning-based Spectrum Sharing and Spatial Reuse in mm-wave Ultra Dense Networks, IEEE Trans. on Vehicular Technology.

[3] Voigt C & Bright J. 2016. The Lightweight Smart City and Biases in Repurposed Big Data. In: Proceedings of HUSO, The Second International Conference on Human and Social Analytics

[4] Bright J, De Sabbata S, Lee S. 2017. Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks? Forthcoming in: *GeoJournal*

[5] McNeill G, Bright J & Hale S. 2016. Estimating Local Commuting Patterns from Geolocated Twitter Data. arXiv preprint arXiv:1612.01785

[6] Quiring E, Arp D & Rieck K. 2017. Fracternal Twins: Unifying Attacks on Machine Learning and Digital Watermarking. arXiv preprint arXiv:1703:05561

*The Alan Turing Institute Programme for Data Centric Engineering would be pleased to provide further detail of any of the issues raised above, either in writing or by way of oral evidence. This response was initiated by Professor Mark Girolami (mgirolami@turing.ac.uk) and coordinated by Dr Weisi Guo and Darren Grey with additional contributions from Dr Din-Houn Lau, Dr Ricardo Silva, Dr. Stephen Victoria, Dr. Tim Sullivan, Dr. Sebastian Vollmer, Prof. Daniel Graham, Dr. Scott Hale, Dr. Mohammed Elshafie, and Professor Julie McCann.*