Data and innovation: the case for experimentation

Frank Kelly

Professor Frank Kelly FRS is Chief Scientific Adviser at the Department for Transport, and Professor of the Mathematics of Systems at Cambridge University. He recently spent a year as visiting professor at the Graduate School of Business at Stanford University. His research interests focus on methods of self-regulation for large-scale complex systems, and he has received many awards for his work on networks.

Government collects and uses a wide range of data to both inform and deliver its policies. This data is generally used for specific purposes and is rarely made easily accessible for other uses. Yet data held by government for one purpose can offer immense benefits in the delivery of other services, particularly when combined or 'mashed' with data from other sources.

Advances in information and communication technologies and the development of more sophisticated and easy-to-use software tools continue to remove the technical barriers to realising data mashing applications. The ability to produce innovative data applications is no longer the preserve of computer scientists, whose role is increasingly in providing the tools and services permitting others to develop highly personalised and specific applications. As a result a diverse community from the public, private and voluntary sectors are engaged in the development of novel data mashing applications.

No single data collector or user, government departments included, can reliably predict how data may be used when combined with data from other sources. Realising these benefits, therefore, requires permitting greater access to data in order to permit experimentation. Among the obstacles to improving access are regulatory and administrative barriers, poor incentives and limited awareness and expertise.

The challenge of realising new data applications is not unique to the public sector. Within the private sector there has been a trend away from a highly controlled development from concept to finished product, towards a more iterative approach where the rapid development of beta version products is followed by testing and further modification of concept and design[1]. The engagement of a diverse stakeholder community during conception, development and testing is essential to success. Such an approach allows the gradual evolution of a product shaped by the stakeholder community and helps identify unforeseen applications for data.

[1] See for example, labs.google.com

The Department for Transport (DfT) has had to grapple with several of these issues, as real-time and archived data have become more and more important for the operation, planning and use of transport networks. Some early examples are instructive.

My first example is MIDAS (Motorway Incident Detection and Automatic Signalling). The most heavily used parts of the motorway network have loop detectors every 500 metres: these sense the presence and speed of a vehicle. A real-time control loop uses this data to adjust a variable speed limit – the aim being to reduce accidents caused by cars approaching a jam too quickly. The data have been archived since 1997. The archive provides insights into complex system behaviour, such as flow-breakdown, where, above a certain density of vehicles, flow becomes unstable[2]. These insights have implications for controlling access to motorways, and for public policy in this area. A further recent application of MIDAS has been to the monitoring of the DfT's Public Service Agreement target for reliable journeys, where the data is mashed with data from TrafficMaster, from GPS traces and from automatic number plate recognition cameras to produce estimates of congestion on the strategic road network[3]. Thus data generated for one purpose has later had uses that could not have been originally envisaged.

My next example is Transport Direct, which aims to provide a comprehensive, easy-to-use, multi-modal, travel information and ticketing service. Transport Direct is, in reality, an enormous virtual team incorporating hundreds of organisations and individuals[4]. Each of over a hundred sources needed to agree to provide their data and also to make it available in common standards and formats. For the first time a unique number was agreed for each bus stop in the country. Such standardisation is essential if data is to be accessible and useful. And the data has many further applications[5]. In particular, the DfT has funded the development of: Accession[6] software, which can bring together bus data from Transport Direct, GIS data from the Ordnance Survey and census data from the Office of National Statistics to help local authorities identify whether people can get to jobs, education, health and other key activities; and mySociety's travel-time maps[7], which can help people make decisions on where to work or live in ways that are better informed by public transport options.

Many drivers use GPS devices to warn of the presence of speed cameras. Should the DfT make a database of road speed limits freely available in open format? Such a database

---

[2] R. Gibbens and W. Werft, Data gold mining: MIDAS and journey time predictors. Significance, volume 2 (2005), issue 3.

[3] Congestion on the Strategic Road Network. DfT: www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transstats_611154.hcsp

[4] Nick Illsley, Chief Executive, Transport Direct: www.egovmonitor.com/node/156

[5] Indeed it is possible to envisage competition in innovative front-ends – see, for example, www.busmonster.com and www.google.com/transit

[6] www.accessiongis.com

[7] www.mysociety.org/2006/travel-time-maps

could encourage the development by equipment providers of innovative in-car systems to inform the driver of the applicable speed limit, with helpful safety implications. It would not necessary for the DfT to invent the in-car systems, but only to be aware that there are several promising and innovative technological approaches. The state encourages market demand, of course, by its enforcement of speed limits. The barriers to the development of an open database are primarily legal, commercial and institutional. Local authorities, which set speed limits and contribute the data, are reluctant to pay fees to the Ordnance Survey for access to the data[8]. The Ordnance Survey is reluctant to allow in-car system providers access to the data unbundled from their other products. There are real issues here about the way we fund data collection and management. The Ordnance Survey and the Met Office have very good data, partly because they have revenue streams that allow them to do their jobs properly. But high data prices can prevent innovation, and Government has several potentially conflicting roles [9].

We are acquiring many new sources of data in transport. Nearly three million journeys a day in London are made by Oystercard, and mobile phone trajectories are potentially a further source. Such data on origin/destination paths and route choice is important for the planning of new transport infrastructure and for evaluating the impact of pricing policies. In a world of increasing incomes and higher transport environmental costs we should expect innovation in the allocation of scarce capacity - an early example is web-based airline booking systems, where prices are adjusted dynamically based on historical and current booking patterns. In the future we can expect the technologies of wireless communication, mobile computing and geographical positioning to transform the economics of transport, allowing better matching of supply to demand with unprecedented levels of precision and speed –an early illustration of the integration of these technologies is pay-as-you-drive insurance[10]. The future will be shaped by the use we make of new technologies, which both resolve some barriers to data access (e.g. better security technologies, access controls, search algorithms) and raise new concerns about data use (e.g. loss of privacy through greater integration of data, data misuse).

The Department for Transport is leading work for the Science and Innovation Ministerial Committee's *Data Grand Challenge* on realising the benefits of data within and outside of government. We are experimenting with more flexible ways of working, particularly in terms of the commissioning and management of projects. We are not seeking to define final data applications but to allow experimentation and the gradual evolution of applications, and we recognize the potential added value of suitably anonymised official data being made available for mashing with other data sources. An early example of this approach is the mySociety travel-time maps mentioned earlier, and this example is helping clarify several of the obstacles to improved access to data. More ambitiously, the challenge hopes to explore how government can move from its current 'control and

[8] The comments of Professor David Rhind, formerly Chief Executive of the Ordnance Survey are especially interesting: D. Rhind, Geographic information and traffic management. FST Journal, volume 18 (2005), issue 9. www.foundation.org.uk/pdf18/fst18_9.pdf

[9] P. Weiss, Borders in cyberspace: conflicting public sector information policies and their economic impacts. gsdidocs.org/gsdiconf/GSDI-7/papers/TSlpPW.pdf

[10] www.norwichunion.com/pay-as-you-drive

communicate' model of data provision, to a more decentralised model in which any unanticipated but legitimate user can find, access and use data. It is widely understood that sensitive personal data, and the definition of a legitimate user, are areas where regulation and institutions find it difficult to adapt to rapidly evolving technology. Examples from this talk show that even where there are no privacy issues, and where public policy objectives are served by wide propagation of information, there are formidable barriers to data access.