Inside the 'Big Data' Factory: Lessons from Health Informatics

Jessica Wardlaw
University College London
Gower St, London, WC1E 6BT
j.wardlaw@ucl.ac.uk

Mordechai (Muki) Haklay
University College London
Gower St, London, WC1E 6BT
m.haklay@ucl.ac.uk

Anna L. Cox
University College London
Gower St, London, WC1E 6BT
anna.cox@ucl.ac.uk

ABSTRACT
This paper highlights wider engineering challenges encountered in the design and development of interfaces for 'Big Data' analysis tools with a small number of users in specialist application domains. Based on six years' collaboration with a UK-based health informatics company processing 16 million hospital admissions records a month, we report challenges associated with commercial pressures, rapid changes in the business environments, and variation in users' computer literacy, functional requirements, technical resources and knowledge of the complex underlying data. Whilst the literature advocates new interaction and analysis techniques for Big Data visualizations in these contexts, this paper cautions that some users cannot spend the time to learn how to interact with them or are constrained by their working environment. Despite these challenges, we show that User-Centred Design and research 'in the wild' can go some way to address these engineering challenges and support the implementation of Big Data business-to-business applications.

Author Keywords
Big data; health informatics; user-centred design

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

BIG DATA, BIG BUSINESS

The volume, velocity and variety of data being created and analysed is ever increasing [2]. IBM state that 90% of the world's data was created in the last two years and 2.5 million gigabytes of data is created daily – enough to fill 27,000 iPads per minute - a phenomenon known as 'Big Data'. Data has been compared to the "oil" of the 21st century [4], but, unlike oil, it will not run out. As mobile access to computers and the internet increases, new and emerging technologies will fuel the growth of data.

Big Data presents significant opportunities and is already transforming business sectors across the global economy: from increased transparency and accountability with open data, to new scientific discoveries, and market-changing products and services, which can be developed using modelling, data analytics and data-driven science. Its volume and variety, and the speed at which it is created and processed, however, can pose huge challenges.

Designing systems to cope with large volumes of data is not a trivial design problem, especially when the users do not have the requisite skills to analyse and visualise such large datasets. A shortage of skilled workers in the data analytics market is cited as one of the key barriers to further data analytics activity [2]; these challenges have hindered the realisation of the opportunities presented by Big Data [10]. User-Centred Design (UCD) is important in facing these challenges because consumers, businesses and academia require usable analytical tools to profit from Big Data; UCD can help to reverse this through the design of tools that effectively support workers to analyse and present an increasingly large volume and variety of data.

While much of the discussion surrounding Big Data applications considers large numbers of users, it is vital to understand the social context. In the 'Social Life of Information', Brown and Duguid remind us that, by engaging the social context in which technology is embedded, better designs and uses will emerge [1]. Technology is most successful where it augments the natural abilities of humans; for example, people identify others as individuals and interact with them in context in a way that computers are unable to replicate.

This position paper will further this idea by discussing the wider engineering challenges for the design and development of interfaces for Big Data analysis tools with a small number of users in specialist application domains. The lead author reflects on a six-year collaboration with Dr Foster Intelligence (DFI), a public-private partnership in between England's National Health Service (NHS) Information Centre and Dr Foster Ltd, who were pioneers in the UK health informatics sector. They provide independent health and social care information to healthcare managers and clinicians to improve clinical effectiveness and efficiency. DFI has produced a range of web-based data analysis tools which give NHS managers access to the Hospital Episodes Statistics (HES) database that contains admitted patient care data from 1989 onwards and outpatient attendance data from 2003 onwards. Whilst live access to 825 million hospital records presents many challenges, with users varying in computer literacy, functional requirements, technical resources and knowledge of the complex underlying data, it offers opportunities that are unavailable in any other country's health system.

Designing effective user interfaces for Big Data is a challenge that user-centred design and evaluation methods (UCM) from the field of Usability Engineering can overcome. The lead author collaborated with DFI to adapt UCM, which revealed the business realities of developing and marketing Big Data applications; they can be a high risk investment for small-medium enterprises. Competitive pressure to keep up with technological advances can tempt companies towards technology-led development processes.

This experience provides a unique perspective on the engineering-related challenges associated with the design of interfaces for Big Data applications and the methodological support required in a specific application domain. This will be informative for other enterprises looking to exploit the growing business-to-business market for Big Data.

INSIGHTS FROM INSIDE A BIG DATA BUBBLE

The major challenge for DFI is to process 16 million records a month, add value to them and enable users to run meaningful analysis; learners and novices need support to achieve their tasks and goals in a timely manner. Many different user types perform different roles (for example, Information Analysts, Clinicians, Public Health Analysts, Medical Directors) and correspondingly different tasks. Users therefore have a wide disciplinary background and vary in computer literacy, often correlated with how long they have been working in the NHS. (For example, users who have had a long career in the NHS can be accustomed to particular working practices and averse to change.) Furthermore, users are spread across England in hospitals that operate in subtly different ways. These user characteristics are not unique to health informatics and can be seen in other emerging Big Data application domains. These include financial services, insurance industry, local government, social services, urban planning, large infrastructure projects, environmental management and climatology, in which the end-users synthesize Big Data to make decisions that impact upon people's everyday lives.

Integrity of the data, transparency in the methodologies used and secure data processing are also important to DFI's customers. Whilst users require design and evaluation support to make sense of it and complete their tasks, they are knowledgeable about the highly complex information that the application processes. A Hospital Standardised Mortality Ratio (HSMR), and the methodology for its calculation, comprised the basis for the formation of the original Dr Foster Ltd, which hospitals use to benchmark their death rate; the organization also adds various flags to each admission record, including readmission within 30 days and patients' demographic and lifestyle characteristics. DFI store the data in an area of the office that requires security access and is protected computationally because of its personal and sensitive nature; the system displays an error message if the user runs a query with few results, which would potentially allow patients to be identified.

Finally there are more systemic, technological challenges. A screenshot study to inform personas development revealed that users integrate Big Data within their local working environment. The screenshot in Figure 1 shows Microsoft Word and Excel open simultaneously; this corroborated interview data that revealed many users rarely use anything other than Microsoft Office applications for their work. Users integrate the data and visualizations into Microsoft Word to create reports, Excel for analysis and PowerPoint to present at meetings. Such information must inform the design of the user interface and visualizations. They are familiar to certain icon and button conventions from these applications and find it difficult to adjust to others. The implications of this are that whilst the latest literature tends to endorse and advocate more current, novel and advanced visualizations [3,5,7], many users are not accustomed to the visualization techniques used and do not have the time to learn them. DFI's users are further limited by the technology available to them in a hospital. The technology infrastructure in NHS hospitals necessitates that tools developed by DFI are compatible with all web browsers; users rarely have a good network connection or the authority to download and use the latest browsers, or any additional software required to use some of the more advanced visual analytics tools. In subsequent interviews for the personas study, one user reported that they logged on in the evening with a glass of wine because their home network connection is better.

USER-CENTRED DESIGN OF METHODS AND TOOLS

A mixture of technical and organizational obstacles prevented the implementation of many recommendations that resulted from studies that the lead author carried out.

Commercial pressures, and the rapidly changing business environments in which Big Data applications are developed, create challenges for their development; many UCM demand more resources than commercial pressures permit Big Data enterprises to use. Whilst the authors'

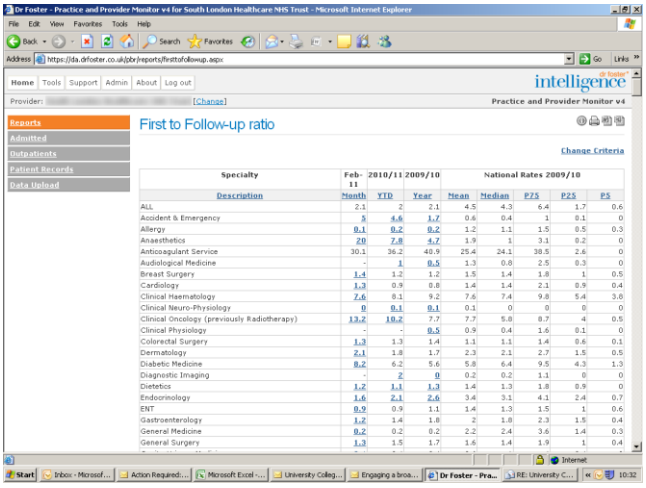

Figure 1 User-Generated Screen Capture

collaboration with DFI had a positive impact in promoting UCD within the organizational culture, the lead author was given limited contact with end users to carry out user testing for commercial reasons; the organization preferred that users did not see unfinished products so that they would not be discouraged to renew contracts and switch to the competition. User testing was eventually carried out in two phases late in the development process, when the organization was confident that the application was sufficiently functional and presentable; the resultant time pressure hampered recruitment of participants (six in phase one, nine in phase two) and implementation of their recommendations. The limited time available to participants, and their geographical spread, further delayed the scheduling of test sessions. Therefore enhanced 'Discount' UCM were used and adapted to the context [6].

Methods included user testing and an online questionnaire to explore the users' level of understanding of maps. Personas were also developed with database server log file analysis, a user-generated screen capture study and interviews during which participants demonstrated a typical task, to provide the developers with a better understanding of the users' goals, behaviours, attitudes, skills and working environment. The authors adapted the way they recruited participants, selected tasks, identified usability problems and reported results according to the business and project requirements [8]. This methodological approach offers unique insights into the challenges of developing usable Big Data applications and facilitates the co-design of new approaches to develop suitable interaction techniques.

Figure 2 The Office "Fishbowl"

Many widely-used UCM were not designed for complex systems such as Big Data applications and are therefore not necessarily appropriate. For the first phase of the collaboration the lead author managed the design and development of geographical analysis functionality for a tool for public health professionals. Geospatial tools have inherent difficulties for users that are not necessarily revealed using discounted UCM [9]. For users to exploit these datasets to their full potential therefore requires new approaches to their design and evaluation.

In addition, organizational and physical aspects of the work environment played an important role. For example, a glass wall (Figure 2) was required because of the data's sensitivity; only a few developers were permitted access to it. Developers likened working behind the wall to working inside a "fishbowl"; it created a 'silo' effect and impeded communication between developers and their colleagues, including designers and customer-facing teams.

CONCLUSION

This position paper demonstrates the need for design and interaction techniques for small-medium data-focused enterprises that develop specialist Big Data applications for a limited number of users who vary in computer literacy, functional requirements, technical resources and knowledge of the complex underlying data. A finite market of customers and high commercial stakes demands the development and adoption of ostensibly trivial and facile interaction techniques and UCD methods, which can be overlooked in the pursuit of more advanced, viscerally impressive, solutions. It also demonstrates a major engineering challenge facing the designers and developers of Big Data applications: many users do not have the time or technology to be able to learn and use more advanced visualizations. Although research 'in the wild' has elucidated difficulties of embedding UCD in such organizations, it can also support them to adopt UCD to their benefit and tackle the engineering challenges they face in the growing Big Data business-to-business environment.

ACKNOWLEDGMENTS

This work was undertaken for an Engineering Doctorate supported by the Engineering and Physical Sciences Research Council (EPSRC) and DFI, and Industrial Fellowship from the Royal Commission for the Exhibition of 1851. We also thank participants in the studies described.

REFERENCES

1. Brown, J.S. and Duguid, P. *The social life of information*. Harvard Business Press, Boston, MA, USA, 2002.

2. Department for Business, Innovation & Skills. *Seizing the data opportunity: A strategy for UK data capability*. London, UK, 2013.

3. Fisher, D., DeLine, R., Czerwinski, M., and Drucker, S. Interactions with big data analytics. *interactions 19*, 3 (2012), 50.

4. Gartner. Gartner Says Worldwide Enterprise IT Spending to Reach $2.7 Trillion in 2012. 2011. http://www.gartner.com/newsroom/id/1824919.

5. Grammel, L. How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 943–952.

6. Nielsen, J. Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. In R. Bias and D. Mayhew, eds., Academic Press, Boston, MA, USA, 1996, 245–272.

7. CACM Staff. Visualizations Make Big Data Meaningful. *Communications of the ACM (CACM) 57*, 6 (2014), 19–21.

8. Wardlaw, J., Haklay, M., and Cox, A.L. Adaptation of Method-resources Between Projects: A Case Study From a Dynamic and Complex Work Domain. *Proceedings of CHI 2013 Workshop "Made for Sharing: HCI Stories of Transfer, Triumph and Tragedy,"* (2013), 4–9.

9. Wardlaw, J. Principles of Interaction. In M. Haklay, ed., John Wiley & Sons Ltd., 2010, 179–198.

10. Wind-Cowie, M. and Lekhi, R. *The Data Dividend*. London, 2012.