

<b>Institution:</b> Aston University
<b>Unit of Assessment:</b> 11: Computer Science and Informatics
<b>Title of case study:</b> Impact of Machine-Learning based Visual Analytics
<p><b>1. Summary of the impact</b></p> <p>Visual analytics is a powerful method for understanding large and complex datasets that makes information accessible to non-statistically trained users. The Non-linearity and Complexity Research Group (NCRG) developed several fundamental algorithms and brought them to users by developing interactive software tools (e.g. Netlab pattern analysis toolbox in 2002 (more than 40,000 downloads), Data Visualisation and Modelling System (DVMS) in 2012).</p> <p><b>Industrial products.</b> These software tools are used by industrial partners (Pfizer, Dstl) in their business activities. The algorithms have been integrated into a commercial tool (p:IGI) used in geochemical analysis for oil and gas exploration with a 60% share of the worldwide market.</p> <p><b>Improving business performance.</b> As an enabling technology, visual analytics has played an important role in the data analysis that has led to the development of new products, such as the Body Volume Index, and the enhancement of existing products (Wheelright: automated vehicle tyre pressure measurement).</p> <p><b>Impact on practitioners.</b> The software is used to educate and train skilled people internationally in more than 6 different institutions and is also used by finance professionals.</p>
<p><b>2. Underpinning research</b></p> <p>The extraction and visualisation of information from complex datasets – such as those found in industry and engineering, the health sector and biological research – is an important area of research internationally. The NCRG at Aston has become one of the leading international research groups in machine learning and its application to data analysis and visualisation.</p> <p>The NCRG has a strong research record in pattern analysis and machine learning, especially for feature extraction and visualisation. The group's research has included major advances in theory and in the developments of new algorithms. For example, researchers in the NCRG invented and developed a number of important data visualisation and density models that are exploited around the world. GTM (Generative Topographic Mapping) [3.1] was developed by three Aston staff: <i>Bishop</i> (Professor 1993-7), <i>Svensen</i> (PhD/postdoc 1995-8) and <i>Williams</i> (Lecturer 1995-8) in 1996. NeuroScale [3.2, 3.3] was developed by two Aston staff: <i>Lowe</i> (Professor 1993-present) and <i>Tipping</i> (postdoc). GTM was developed into a hierarchical model in 2002 [3.5] (<i>Tino</i> postdoc 2000-3) and <i>Nabney</i> (then Senior Lecturer: at Aston 1995-present). These models project data non-linearly from a high-dimensional space to a lower (usually two-dimensional) space where it can be plotted. The algorithms have been shown to provide a significant improvement in performance and interpretability over conventional algorithms, such as Principal Component Analysis. The GTM provides a probabilistic model of the data, while Neuroscale is topographic (distance-preserving).</p> <p>Building on these concepts, the NCRG has established a unique capability through the development of new algorithms and the creation of software tools that implement those algorithms. This work came together in 2002 in the development of Netlab, a toolbox of open-source Matlab pattern analysis software that provides a platform for further research, application development, and technology transfer. The accompanying textbook [3.4] has been very influential. Netlab contains a wide range of pattern analysis algorithms, including GTM and Neuroscale, and is freely available from (<a href="http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/">www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/</a>).</p> <p>Over the years, the group has contributed both to advances in the theoretical underpinnings for data visualisation and has also carried out research into the practice of visualisation, especially for non-statistical users [3.6]. This work led to the use of the Netlab toolkit to build applications in a wide range of domains. Drug discovery and microarray analysis provided the initial proving grounds for these visualisation technologies. A generic data visualisation tool, DVMS (Data Visualisation and Modelling System) was developed with Pfizer in 2006 and made freely available on the NCRG website (see link above) since March 2012. This tool brings together the novel data projection algorithms developed in the NCRG with a rich interface that allows the non-statistically trained user to interactively interrogate and explain features of the dataset.</p>

### 3. References to the research

All citation counts are taken from Google Scholar on 15<sup>th</sup> October 2013. The three references that best indicate the quality of the underpinning research are indicated with a triple asterisk \*\*\*.

- 3.1. \*\*\* Bishop, C. M., Svensen, M. and Williams, C. K. I., GTM: The Generative Topographic Mapping. *Neural Computation* 10 (1), 215-235, 1996. doi:10.1162/089976698300017953. Evidence for quality: 1009 citations, of which fewer than 20 are self-citations or from Aston authors. It is still being cited more than 60 times per year. Journal (5-year impact factor of 2.5) is in the top handful in the field.
- 3.2. \*\*\* Lowe, D. and Tipping, M. E., Feed-forward neural networks and topographic mappings for exploratory data analysis. *Neural Computing and Applications* 4, 83-95, 1996. doi:10.1007/BF01413744. Evidence for quality: 120 citations. This application-oriented journal was chosen for its potential to influence practitioners.
- 3.3. Tipping, M.E. and Lowe, D., Shadow Targets: A Novel Algorithm for Topographic Projections by Radial Basis Functions, *Neurocomputing*, 19, 211-222, 1997. doi:10.1016/S0925-2312(97)00066-0. Evidence for quality: 87 citations for this paper and its conference version.
- 3.4. Nabney, I. T., *Netlab – Algorithms for Pattern Recognition*. Springer Verlag, London, 2002. Software available from <http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/> ISBN: 1852334401. Evidence for quality: text book has been through three reprints, has been cited 918 times, and is still being cited more than 50 times per year. Software has had more than 40,000 downloads since first being made available.
- 3.5. \*\*\* Tino, P. and Nabney, I. T., Hierarchical GTM: Constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 639-656, 2002. doi:10.1109/34.1000238. Evidence for quality: 68 citations. This is a top international journal in the machine-learning field (5-year impact factor more than 6) with peer review standards to match.
- 3.6. Maniyar, D. M. and Nabney, I. T., Visual data mining using principled projection algorithms and information visualization techniques, *ACM SIGKDD*, 643-648, 2006. doi:10.1145/1150402.1150481 Evidence for quality: 8 citations. Won a best student paper prize at the conference.

Grants collaborative with industry in the census period that have developed and made use of visualisation techniques based on the research above: these are evidence of the quality of the underpinning research and continued activity of the Unit in this domain.

- A. Nabney, EPSRC CASE with IGI Ltd. (£81k) 2007-2010. "Visualisation of geochemical data"
- B. Nabney, BBSRC CASE with Pfizer Central Research, (£78k) 2007-2011. "Improved *in silico* prediction of molecular properties"
- C. Nabney, EPSRC CASE with AgustaWestland (£82k) 2011 - present. "Diagnostics and inference for helicopter vibration data"
- D. Nabney, Centre for Defence Enterprise with Daden (£32k) 2011. "Use of virtual immersive worlds for data visualisation"
- E. Nabney, short KTP with Lein Diagnostics (£20k) 2012. "Improved glucose monitoring for diabetics"
- F. Nabney, WheelRight Ltd. (£34k) 2012 – present. "Pattern analysis to improve in-road measurement of tyre pressure"
- G. Lowe, Centre for Defence Enterprise (Dstl), (£24k) 2011. "Multi-Source Intelligence: Challenge 2: Probabilistic Visualisation Maps"
- H. Lowe, EPSRC Industrial CASE with Thales (£90k) 2011 - present. "Topographic Information Visualisation"

### 4. Details of the impact

#### Process

Visual analytics is a powerful and generic technology to provide insight into data: thus we have sought to create impact in a wide range of domains. The process for creating impact started with the implementation of a well-engineered software toolbox (Netlab) that is reusable, can be deployed quickly, and forms the underpinning technology for larger systems. Netlab supported the creation of user-centred tools DVMS (originally for Pfizer), milva (gene expression array analysis), et al. DVMS in particular has been a key step in creating impact since it has enabled researchers

at Aston to deploy visualisation methods on a wide range of projects (subsections 4.2 and 4.3) and work directly with end users (who do not have statistical training) to understand the data and thus improve the accuracy of data analysis. Our standard contracts allow the reuse of generated IP in non-competing domains, which has enabled us to build the capability of our visual analytics tools across multiple projects for different companies. Thus high-quality data visualisation is a fundamental part of the data modelling process on many industrial projects carried out by the NCRG.

The dissemination strategy has had impact at its heart from the beginning. Papers were published in both scientific and application journals; inter-disciplinary conferences (e.g. SIGKDD) and user/industrial workshops (e.g. Natural Computing Applications Forum, Chemometrics SIG) have been targeted; publicly available open-source software and industrial training courses (designed in collaboration with industry) increased the impact on practitioners.

Applications where visual analytics is particularly useful can be characterised by large quantities of uncertain information where there is a need for non-statistically trained users to interpret data and understand its analysis, and so organisations with this sort of data have been targeted.

Relationships with companies have been formed through personal networks (e.g. Pfizer, IGI, Thales) and contacts made with Aston's Business Partnership Unit (e.g. WheelRight, Daden, AgustaWestland, Lein): a mixture of special events targeted at companies with appropriate data and company-led interactions (often inspired by the high visibility of the NCRG and Netlab).

### **1. Industrial Products**

A number of companies have incorporated the data visualisation work of the NCRG directly into their products.

- Pfizer Central Research has supported our work since 2000 through funding (2 directly funded PhDs and 1 CASE studentship; BBSRC funded a postdoc) the development of practical visualisation algorithms and methods [5.1] (and has collaborated in writing joint papers). This culminated in the development (2008) of an interactive visualisation tool used by Pfizer's chemists and biologists (rather than statisticians) to interpret and analyse screening results (e.g. biological activity, toxicity etc).
- Integrated Geochemical Interpretation Ltd, a petroleum geochemistry consultancy company which operates world-wide, sells p:IGI, a software product for oil and gas exploration geochemistry with a 60% share of the worldwide market. IGI co-funded a PhD CASE student [5.2] who, after completing his thesis in 2010, worked for the company to implement our visualisation algorithms to enhance the p:IGI tool [5.7]. The company started a £500k TSB grant in 2013 (jointly with Aston and Daden, see also below) to drive further development of this product and enable them to expand into new business sectors.
- Intelligence data visualisation: A prototype information visualisation system was delivered to Dstl (Defence Science and Technology Laboratory) in 2011 under a contract [3.G, 5.4] on Collaborative Multi-source Intelligence related to integrating spatio-temporal and network analysis incorporating measures of uncertainty.
- Thales is now exploring the use of topographic information visualisation using very high-dimensional submarine sonar array data as part of a collaborative industrial CASE project. The idea is to synthesise the vast amounts of information into a form suitable for human interpretation by skilled operators on board submarines.

### **2. Improving business performance**

Visual analytics and the DVMS software have played a fundamental role in collaborative industrial research projects in order to understand data, detect outliers and select important features. This improved analysis has led to the development of new and improved products.

We worked (2010-1) with Daden, an SME based at Birmingham Science Park Aston, on a £35k contract for Dstl (the Cyber and Influence Centre, part of the Centre for Defence Enterprise) for immersive data visualisation and its application to defence intelligence analysis. The results of this research formed part of the case for a significant corporate investment in Daden from BAE Systems and the development of the Datascape product [5.6] in 2012-3.

Nabney consulted for Select Research Ltd. [5.3] to develop the data visualisation and analysis for the Body Volume Index (BVI), a novel obesity metric replacing BMI, based on volumetric measures

with a white-light scanner. This metric is much better correlated with biomarkers of vascular disease than BMI. Select launched BVI in November 2010 and have been in trials for the NHS.

In a CASE studentship with AgustaWestland [3.C] to develop a condition monitoring system for helicopter airframes, DVMS has been used to select frequency bands and sensors that will provide the best detection of faults. The visual nature of the results has enabled us to explain the selection process to the engineers and demonstrate the transitions between different flight modes. The results of this are a prototype for detecting the need for structural maintenance on the airframe.

WheelRight [5.5] makes sensors for measuring vehicle tyre pressure which are embedded in the road so that measurement is performed simply by driving over the sensors and the results are sent by text to the driver. Correct tyre pressure maintenance is done poorly and irregularly.

WheelRight's goal is to make this process automatic, specific, accurate and easy to use. A study by Bridgestone Tyres estimates tyre under-inflation is costing the EU €2.8 billion each year in wasted fuel and adding 4.8 million tons of CO<sub>2</sub> emissions. This high cost is due to drivers' unwillingness to use the current tools. There is now a better way to radically reduce this waste.

In a directly-funded research contract (2012-3) we have significantly improved the results from Wheelright's vehicle tyre pressure measurement instrument. We used visualisation to understand the complex data signals and then developed new predictive models that significantly improved the measurement accuracy, particularly on HGVs and buses, so that it now meets the requirements of their customers. WheelRight have installed three complete systems used each day by a range of vehicle types. At a municipal bus depot with a fleet of 80 buses over 2,000 tyre pressures are taken each week, at a HGV fleet operator this increases to over 14,000 and at the entrance to a technology site over 10,000 are recorded. In total the company has automatically recorded over a million tyre pressures in a twelve-month period in 2012-3 from these three sites alone. Taking this number of manual tyre pressures would be impractical and very expensive; estimated labour costs would exceed £500k p.a. The bus operator is achieving cost savings from lower fuel consumption, longer tyre life and labour exceeding £50,000 p.a. The system has also undergone tests in 2013 by the National Measurement Office to confirm the system's accuracy for all vehicle types.

Building on the analysis of clinical trial data (2012) by Aston researchers using visualisation techniques, Lein Diagnostics has developed an optical measurement device for blood glucose. In 2013, Lein has received further support (£94k) from the NHS to develop an improved version of the meter. The aim is to develop a meter that will achieve the measurement accuracy specified by the ISO and FDA in order to displace the widely disliked finger-stick meters used by diabetics.

### **3. Impact on practitioners**

The primary means of impact on practitioners is the Netlab toolbox. So far there have been more than 40,000 downloads of Netlab by the academic and business users worldwide since its release. Currently, it is averaging 6,200 page-views per year (thus c. 34,000 over the REF census period) and 3,363 downloads per year (thus c. 18,500 over the REF census period).

A portfolio manager at Tudor Capital Europe LLP (who carried out a PhD at Aston in data visualisation) in charge of a large macroeconomics trading portfolio uses data visualisation and dimensionality reduction methods (particularly GTM) as a key tool in his portfolio management strategy. He uses visualisation to interpret the very large number of macroeconomic indicators which often helps him identify new macro trading themes earlier than most market participants.

Several universities also use the Netlab software toolkit and associated book [3.4] in pattern recognition courses, including: Edinburgh; Southampton; Portland State University; the University of California, San Diego; the University of Warsaw; and the Czech Technical University in Prague. This means that training in these data visualisation algorithms is not confined to Aston University, thus increasing the supply of highly-skilled practitioners.

### **5. Sources to corroborate the impact**

- 5.1. Pfizer Central Research: Hub Submissions Manager
- 5.2. IGI Ltd.: Geochemical Consultancy Manager
- 5.3. Select Research Ltd.: Managing Director
- 5.4. Dstl: Fellow in Informatics.
- 5.5. WheelRight: Chief Executive.
- 5.6. Datascape <http://www.daden.co.uk/solutions/datascape/>
- 5.7. p:IGI <http://www.igiltd.com/pigi-3.html>