

Institution: Newcastle University		
Unit of Assessment: 10		
Title of case study: The creation of a data science and machine learning company		
Period when the underpinning research was undertaken: 2001-2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Dr Colin Gillespie	Senior Lecturer	2002-present
Prof Darren Wilkinson	Professor of Stochastic Modelling	1996-present
Prof Richard Boys	Professor of Statistics	1985-2019
Period when the claimed impact occurred: 2016-2020		
Is this case study continued from a case study submitted in 2014? N		
1. Summary of the impact (indicative maximum 100 words) <p>Newcastle University (NU) Mathematical Sciences research has created Jumping Rivers PLC, a Newcastle-based analytics company specialising in data science and machine learning. The research has enabled commercial, economic and education impacts across diverse sectors throughout the UK and overseas. Jumping Rivers has directly met major objectives of the North East Strategic Economic Plan, such as “creating more and better jobs” and providing growth within the digital sector.</p> <p>Specifically, NU research has created impact with corroborated evidence through:</p> <ul style="list-style-type: none"> - establishing a new company; - enabling an annual turnover by 2019/20 of £638K; - achieving a year on year increase in operating profits from £4K in 2016/17 to £184K in 2019/20; - creating 13 jobs; - delivering solutions in sectors such as biotechnology (e.g. Fujifilm Diosynth Biotechnologies) and utilities (Northumbrian Water Limited). 		
2. Underpinning research (indicative maximum 500 words) <p>E-Science Projects</p> <p>Starting in 2001, colleagues in the School of Mathematics and Statistics (now the School, of Mathematics, Statistics and Physics) collaborated with researchers in Computing Science and Biomedical Sciences on several BBSRC E-science and bioinformatics research programmes [G1-G4]. Relevant aspects of two of these projects (BASIS and CaliBayes) are summarised below. The underlying premise of all projects was a commitment to Bayesian modelling (e.g. [P1]) and a recognition of the need to provide complete packages of data science techniques. A common theme was the implementation of computationally intensive methods running on remote computer clusters. This nascent cloud computing pre-dated the rise of the current ubiquitous cloud (Amazon AWS was launched in 2006), enabling collaborative stakeholders to use this technological solution much earlier than what was commercially available at the time.</p>		

The BASIS (Biology of Ageing e-Science Integration and Simulation) project [P1] was a flexible web-service based modelling system which was developed to allow parallel and distributed approaches to simulation and inference [P2]. It was effectively a set of web services that allowed users to incorporate simulation tools into their workflows [P3]. In 2002, web-services were very much in their infancy and connecting disparate services was challenging. Today, many Jumping Rivers (JR) data science projects consume data from a variety of application programming interface (API) endpoints, e.g. weather data. The key to combining different data sources is a standard interface. The BASIS project provided the foundation for these skills by developing the SBML framework that enables the exchange of statistical models and data sets [P4].

The second project, CaliBayes, developed techniques for calibrating complex stochastic models against experimental data, with careful attention to the selection of parameter combinations at which to run expensive and time-consuming simulators [P5]. Crucially, these techniques were running statistical algorithms in the pre-cloud. Indeed, figures 1 & 2 in [P5] provide a data pipeline architecture, that includes API requests, load balancing and secure data access points, that underpin the architecture currently used by JR for data pipelines within their industrial and business client base.

Creation of a data analytics company

The primary aim of research in statistical methodology and data science is to produce new or improved techniques for use in practical applications. While academics may be able to install and experiment with the latest statistical software, commercial/enterprise users face significant challenges in implementing new techniques. These hurdles include accessing data, maintaining data security, upgrading statistical software, selecting the methods, applying model diagnostics, and moving from a testing to a production environment.

Jumping Rivers PLC is an independent company that was formed in 2016 because of the NU research and alongside an increased demand for support in implementing data science and statistical methods. The Newcastle research [P1-P5] and programmes [G1-G4] provided early recognition that science could best leverage increased computing power through integrated approaches that cover all aspects of the data pipeline, from data entry and storage through Bayesian modelling and computationally intensive statistical inference on remote computer clusters, to visualisation and prediction.

The research focussed on applications in systems biology and the biology of ageing, but the techniques are generic. During the research and subsequently, it became clear that diverse organisations needed integrated data science approaches. With the School's core focus being research and teaching, an opportunity was identified to commercialise the research. Gillespie, who had been a research associate on one of the e-science grants and was subsequently a Lecturer then Senior Lecturer in the School of Mathematics and Statistics, was frequently consulted from around 2010 by external organisations seeking to implement their own circle-of-data systems. Recognising that demand was high and likely to increase, Gillespie reduced his University employment to part-time in 2016 and co-founded JR.

3. References to the research (indicative maximum of six references)

Publications

[P1] Bayesian inference for a discretely observed stochastic kinetic model: Boys RJ, Wilkinson DJ & Kirkwood TBL (2008). *Statistics and Computing* 18, 125-135 <https://doi.org/10.1007/s11222-007-9043-x>

[P2] Towards an e-biology of ageing: integrating theory and data: Kirkwood TBL, Boys RJ, Gillespie CS, Proctor CJ, Shanley DP, & Wilkinson DJ (2003). *Nature Reviews Molecular Cell Biology* 4, 43–249. <https://doi.org/10.1038/nrm1051>

[P3] Tools for the SBML Community: Gillespie CS, Wilkinson DJ, Proctor CJ, Shanley DP, Boys RJ, Kirkwood TBL (2006). *Bioinformatics* 22, 628–629. <https://doi.org/10.1093/bioinformatics/btk042>

[P4] Systems Biology Markup Language (SBML) Level 3 Package: Distributions, Version 1, Release 1: Smith LP, Moodie SL, Bergmann FT, Gillespie CS, Keating SM, Konig M, Myers CJ, Swat MJ, Wilkinson DJ, Hucka M (2020). *Journal of Integrative Bioinformatics* 17, 20200018 <https://doi.org/10.1515/jib-2020-0018>

[P5] CaliBayes and BASIS: integrated tools for the calibration, simulation and storage of biological simulation models: Chen Y, Lawless C, Gillespie CS, Wu J, Boys RJ, Wilkinson DJ (2010). *Briefings in Bioinformatics* 11, 278–289. <https://doi.org/10.1093/bib/bbp072>

Grants

[G1] Bayesian inference for stochastic kinetic genetic regulatory networks. BBSRC, BIO14454, 2001, £120K, Boys, Kirkwood, Wilkinson.

[G2] Biology of Ageing e-Science Integration and Simulation System (BASIS). BBSRC BEP17042, 2002, £442K, Kirkwood, Boys, Proctor, Wilkinson.

[G3] Integration of Grid-based Post-genomic Data Resources Through Bayesian Calibration of Biological Simulators. BBSRC BBS/B/16550, 2005, £352K, Wilkinson, Boys, Kirkwood.

[G4] ComparaGrid - Enabling GRID Technologies for Comparative Genomics. BBSRC BBS/B/17158, 2005, £337K, Wipat, Boys, Pocock, Watson, Wilkinson.

4. Details of the impact (indicative maximum 750 words)

Creation of a data science company

Jumping Rivers is a Newcastle-based analytics company specialising in data science and machine learning [E1]. The company commenced trading in October 2016 and is co-located on the Newcastle Helix site alongside the multi-million-pound National Innovation Centre for Data. JR work impacts on businesses in fields as diverse as biotechnology, finance, DigiTech, manufacturing, chemicals, utilities, and healthcare.

“Newcastle University’s research has been fundamental in creating our company Jumping Rivers, providing the company’s core infrastructure and enabling a diverse range of services to be offered to our customers”. [E2].

Meeting North East challenges

Compared to other parts of the UK, the North East has traditionally experienced fewer and lower quality job opportunities. JR has directly met several objectives of the North East Strategic Economic Plan [E3] such as “creating more and better jobs” and providing growth and services within the digital sector. JR is a highly skilled team with individuals holding qualifications including four PhDs in statistics and six MScs in statistics/computing science. All staff are situated in the North East delivering projects on a global level. The ten technical members of staff specialise in areas such as simulation tools, optimised workflows, cloud-computing, statistical inference, and parallel and distributed approaches.

Impacts

Since the financial year 2016-17, JR [E2, E4]:

- annual turnover has increased from £49K to £638K in 2019/20
- operating profits have increased from £4K to £184K in 2019/20
- staff numbers have increased to a headcount of 13 (12.4 FTE) as of March 2021
- has delivered projects for over 100 organisations in Europe, North America, and Africa
- has provided pro-bono sponsorship for over 40 global events, for example in the Ivory Coast, Uganda, and South Africa

The techniques created through Newcastle University e-science projects have been developed so that JR now offer a complete service through ongoing support, applied research, training, and consulting. Projects include establishing effective data infrastructures, running robust computational statistical models, generating reproducible data analysis reports, and enabling statistical inference. All these approaches enhance business decision making through the improved use of data.

JR has delivered impact across many areas; two exemplars are provided below covering the biotechnology and utilities sectors:

1. Fujifilm Diosynth Biotechnologies (FDB) is a global-based organisation specialising in cell culture, microbial fermentation, and gene therapies. Since 2018, JR has collaborated with FDB to optimise their data collation, usage, and analysis process. FDB has multiple research groups situated in a variety of locations across the world, these teams regularly share large complex data sets. Working closely with a FDB team, JR built a cloud-based data pipeline to conduct process and analytical risk assessments. The risk assessments, built around a Failure Modes and Effect Analysis platform, enable FDB and their clients to identify, numerically score and visualise risks to manufacturing on the journey to commercialising therapeutic drugs.

The digital application is used by all FDB sites globally and has increased their proficiency, elevating client facing capabilities. There are numerous benefits to this work including; more effective prioritisation of tasks in late stage development; reduced chances of errors; enhanced communication of risk to clients; and improved demonstration of risk management systems to global regulatory health authorities including the Food and Drug Administration and the European Medicines Agency. The work has helped FDB to demonstrate their recognition of the importance of effective risk management.

“Jumping Rivers and Newcastle University have both contributed significantly to our global operations”. [E5].

2. Northumbrian Water Limited (NWL) serves 1.3 million properties in the North East of England, and 794,000 properties in Essex and Sussex, primarily through water and sewage services. In recent years, reducing leakage and interruptions to supply have been high priority targets for NWL, its customers, and Ofwat (the water industry regulator).

Since 2018, JR have developed and integrated within NWL data-driven solutions to address leakage and supply interruptions problems. Development of the e-science application was on the Microsoft Azure Cloud, which matched the internal cloud of NWL, and allowed constant feedback to the new application. Data was gathered via APIs from a variety of internal and external sources. Reducing the time taken for 35 engineers to access the required data from two hours to 15 minutes has resulted in annual savings, reduced leakage and reduced interruptions.

The insights provided by JR have been valuable for enhancing NWL service delivery, primarily for NWL customers, but also for NWL themselves, as high performance in these areas is linked to considerable financial incentives.

“Jumping Rivers have demonstrated the value and capabilities of new technologies which we now are looking to adopt further within NWL. Their expertise spans a range of analytics, data science and deployment technologies -including the Microsoft Azure Cloud - which they leverage to build impactful and insightful Proof of Concept models and analyses”. [E6].

NU’s research through JR has provided the bedrock of these case studies and many other projects, enabling commercial, economic, and educational impact within diverse contexts.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[E1] Website for Jumping Rivers. www.jumpingrivers.com Provides evidence of the existence of the company and its services to business and industry.

[E2] Testimonial from the Director of Jumping Rivers. Provides evidence that Newcastle research created the company, financial impacts, and impact types.

[E3] North East Strategic Economic Plan. <https://www.nelep.co.uk/wp-content/uploads/2019/02/executive-summary.pdf> Provides evidence of regional objectives that the establishment of Jumping Rivers has met.

[E4] Jumping Rivers Annual Accounts for the year ended 2020 (page 4). Provides evidence of the annual turnover and operating profits financial impact.

[E5] Testimonial from a Staff Scientist (Statistics and Computational Engineering) at FDB. Provides evidence of the impact of Jumping Rivers and Newcastle University within the biotechnology sector.

[E6] Testimonial from an Intelligence Manager (Intelligence and Analytics) at NWL. Provides evidence of the impact of Jumping Rivers within the utilities sector.