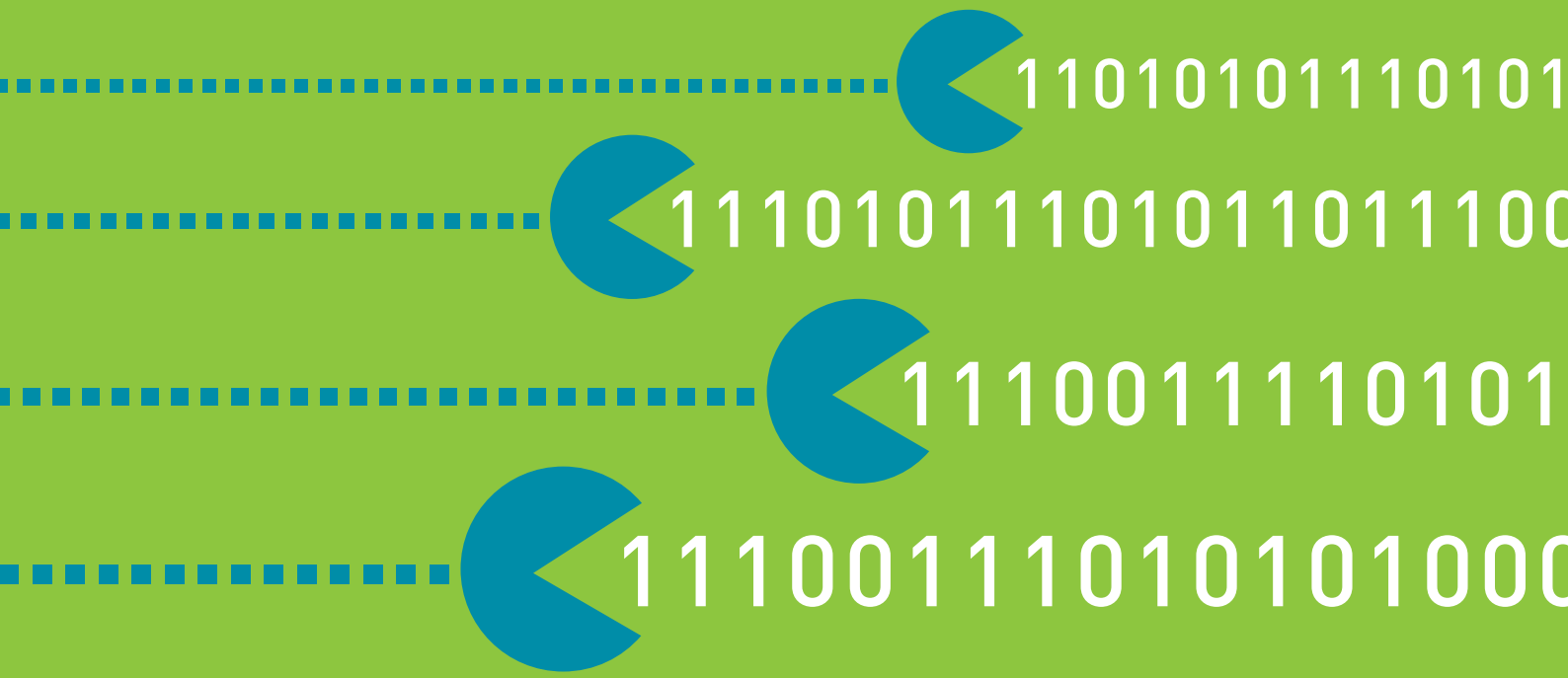
Nesta

SKILLS OF THE DATAVORES

TALENT AND THE DATA REVOLUTION

Juan Mateos-Garcia, Hasan Bakhshi and George Windsor

JULY 2015



Co-Investment from
UKCES
UK COMMISSION FOR EMPLOYMENT AND SKILLS
Employer Investment Fund

creative skillset
Developing world class talent

ACKNOWLEDGEMENTS

Andrew Whitby's contributions were integral to the design of the business survey, and **David Axford** and **Andrew Croll** at Ipsos MORI supported its implementation.

Hetan Shah, **Olivia Varley-Winter** and **Roeland Beerten** at the Royal Statistical Society, and **Kion Ahadi** at Creative Skillset provided valuable feedback throughout the project.

Our ideas about data analytics skills and policy have benefited greatly from discussions with a large number of individuals including **Sir Ian Diamond**, **Samuel Roseveare**, **Rosalind Lowe**, **Professor Patrick Wolfe**, **Professor Sofia Olhede**, **Paul Driver**, **Jane Tory**, **Sandy Grom**, **Steven Bond**, **Keith Dugmore**, **Anjali Samani** and **Kim Nilsson**.

Nesta...

Nesta is an innovation charity with a mission to help people and organisations bring great ideas to life.

We are dedicated to supporting ideas that can help improve all our lives, with activities ranging from early-stage investment to in-depth research and practical programmes.

Nesta is a registered charity in England and Wales with company number 7706036 and charity number 1144091.
Registered as a charity in Scotland number SCO42833. Registered office: 1 Plough Place, London, EC4A 1DE.

www.nesta.org.uk

SKILLS OF THE DATAVORES

TALENT AND THE DATA REVOLUTION

CONTENTS

EXECUTIVE SUMMARY

WHAT ARE THE SKILLS IMPLICATIONS OF THE DATA REVOLUTION?

Digitisation has brought with it bigger volumes and varieties of data, which are being generated at faster velocities. The opportunities to create value from this data are quite simply bewildering. We see this manifested in innovative data-driven products and services, improvements in processes, and more informed decision-making across the economy and society.

However, realising this value requires access to the right skills: data engineering skills to develop a robust data infrastructure, data analysis skills to extract valuable insights from data, and business skills to apply them.

What are those skills exactly? How easy are they to find in the labour market? What are the implications of the data explosion for education and skills policy in the UK?

We have partnered with Creative Skillset to address these questions, focusing on the data analysts charged with transforming data into business insight that drives action. This work builds on Nesta's earlier research on the behaviour of Datavores – our name for those businesses which rely on data for commercial decision-making.[1] The work is also part of *Seizing The Data Opportunity,* the Data Capability Strategy that was published by the Government in 2013, which, together with previously published research by the Tech Partnership and a new report from Universities UK on analytics skills supply, adds up to a substantive body of research on data skills in the UK.

Together this research paints a picture of severe data skills shortages and gaps which should be of deep concern to industry, educators and policymakers. Timely, corrective action is needed across the education and skills system to ensure that UK businesses can fully benefit from the opportunities that data offers. This report is accompanied by a policy briefing jointly developed by Nesta and Universities UK setting out actions to accomplish this.

FINDINGS

The first output of Nesta's data skills research, 2014's *Model Workers*, drew on 45 in-depth interviews to specify the profile of the 'perfect data analyst'. It showed that businesses exploring bigger and messier datasets are looking for analysts – data scientists – with a mix of analytical and computing expertise, domain knowledge, business know-how and communication skills. Our respondents reported that these professionals are very hard to find. The analytical talent coming out of university lack important technical skills; experienced talent is expensive; there is a dearth of individuals with the right mix of skills, and some businesses do not in any case have the internal capability to recruit effectively.

An outcome of this is a widespread perception of a data talent 'crunch' preventing UK businesses from exploiting their data, and in some cases it is even resulting in an offshoring of analytical capability to outside the UK.

In this report, we explore these issues through the experiences of a sample of 404 medium- and large-sized UK businesses in six sectors (Creative Media, Financial Services, Information and Communications Technology (ICT), Manufacturing, Pharmaceuticals and Retail).

Although our sample is not representative of the wider population of UK medium- and large-sized firms, it provides a timely snapshot of the analytical skills situation among companies where data plays an important role in business.

What does it show?

COMPANIES IN OUR SAMPLE ARE USING A VARIETY OF STRATEGIES TO CREATE VALUE FROM THEIR DATA

We have found four distinct 'Data Groups' in our sample, depending on what type of data they work with, and how. Three of them are 'Data-active' in the sense that they are either data-driven (Datavores), work with large volumes of data (Data Builders), or combine data from different streams (Data Mixers). A fourth group, comprising 30 per cent of our sample, work with few, small datasets, and rarely use analysis to make decisions. We refer to them as the 'Dataphobes'.

Analysts in Data-active companies create an impact not just by finding ways to save on costs, but also by seeking out new business opportunities, generating revenues through data products and services, and by improving customer loyalty. By contrast, the Dataphobes appear to have passed up the commercial opportunities emerging form the data revolution.

We find all Data Groups present in all industries, consistent with the idea that data analytics has features of a General Purpose Technology, in that it drives innovation and growth in many different sectors.

DATA-ACTIVE COMPANIES PERFORM BETTER THAN DATAPHOBES

Data-active companies are significantly more likely to self-identify as product/service innovators and process innovators than Dataphobes. For example, Data Builders are over 50 per cent more likely than Dataphobes to say they launch products and services ahead of their competitors.

Our econometric analysis uncovers a positive and significant association between data activity and productivity. This is particularly the case for Datavores and Data Builders, which we show to be, on average, over 10 per cent more productive than Dataphobes, after controlling for other factors.

DATA-ACTIVE COMPANIES ARE RECRUITING LARGE NUMBERS OF ANALYSTS, AND BUILDING A DATA SCIENCE CAPABILITY

They are between two and three times more likely than Dataphobes to have sought to recruit in the previous 12 months. In addition to recruiting experienced talent (from inside and outside their own industries), they are also strongly reliant on talent straight from universities. Around two-thirds of Datavores and Data Builders recruit undergraduates, and more than one-third recruit analysts straight from doctoral programmes.

Data-active companies are also building up their data science capabilities by hiring analysts from numerical disciplines like Mathematics and Statistics at the same time as programmers from Computer Science. They are between two and three times as likely to do this as Dataphobes.

DATA-ACTIVE COMPANIES ARE SUFFERING WORSE SKILLS SHORTAGES

Our findings confirm that there is a crunch in the labour market for analytical talent that is particularly afflicting highly innovative, high-performance Data-active businesses.

Data-active companies are far more likely to report difficulties filling analytical vacancies –for example, as many as two-thirds of Datavores who sought to recruit in the previous 12 months mention problems with at least one vacancy, compared with 40 per cent of Dataphobes. Finding talent with the right domain knowledge, the right mix of skills (e.g. data scientists), experience, and business know-how to apply data in a commercial context is much harder than finding people with the right technical skills (including data manipulation and analysis).

COMPANIES IN OUR SAMPLE ARE UP-SKILLING THEIR WORKFORCE INTENSIVELY, AND INNOVATIVELY

Few companies in our sample think that their analytical workforce is fully equipped with the skills needed to create value inside the business. Consistent with this, they engage in high levels of training activity. Almost 80 per cent of our respondents report providing in-house training to their analysts, and 70 per cent make use of external training.

Substantial numbers of companies (between one- and two-thirds) are involved in online and peer-based learning, reflecting the vibrant ecosystem of hands-on learning, communities of practice and meet-ups in the Big Data/Data Science/tech space. By contrast, universities are the least popular source of training, being used by just one in five companies.

WE SEE SOME SECTORAL DIFFERENCES IN DATA WORK, BUT ALSO CROSSOVER IN DATA APPLICATIONS AND TALENT FLOWS

A company's data sources, analytical methods and areas of application are shaped by the sector in which it operates, something that is reflected in our findings. At the same time, we see many instances of innovative use of data across sectors (e.g. Manufacturers routinely using social media data, and Creative Media companies tapping into Open and Government Data), and of unexpected talent flows across industries. This supports the view that there is great scope for cross-sectoral knowledge sharing and collaboration in this area. Data innovations in one sector are often applicable elsewhere in the economy, and the same is true for data analysis skills.

WHAT'S AT STAKE?

The stakes for the UK economy cannot be understated. If, for example, in our sample of firms all the Dataphobes were to behave like Datavores, our results suggest this would be associated with an overall 3 per cent uplift in productivity. To put this into context, at the macroeconomic level a 3 per cent uplift could, according to OECD statistics,* represent roughly one-fifth of the UK's productivity gap with the rest of the G7.

POLICY IMPLICATIONS

One would be hard pressed to find an industry today where knowledge does not play a central role, or where the ability to learn rapidly does not confer upon its owner a significant competitive advantage.

The data explosion brought about by digitisation is multiplying the opportunities to create knowledge and accelerate learning, and consequently, the opportunities to innovate and grow. Nesta's three-year programme of research in this area has quantified the breadth and scale of data opportunities in the UK economy. It has also demonstrated the importance of developing the human resource and organisational capabilities needed to realise these opportunities – a challenge that is made significantly more difficult by the presence of the skills shortages reported by the Data-active companies we have surveyed, and the apparent ambivalent attitude to data shown by a significant minority of companies – the Dataphobes – we have identified too.

Undoubtedly, data analysis will only gain in importance in UK industry over time and, unless corrective action is taken, we should expect the skills shortages to worsen too.

How can we avoid this gloomy situation, and make sure that UK businesses have access to the analytical talent that they need to thrive in a data-rich world?

To answer this question, we have developed a policy briefing jointly with Universities UK, *Analytic Britain*,** where we set out a major programme of policy recommendations to remove blockages in the pipeline for analytics talent in the UK, spanning schools, universities and the labour market.

*This estimate is based on http://stats.oecd.org/Index.aspx?DataSetCode=PDB_LV

**For more information, see: http://www.nesta.org.uk/publications/analytic-britain

1. INTRODUCTION

THE (BIG) DATA EXPLOSION

The digitisation of social life, industry and commerce has resulted in a data explosion. The numbers involved are extraordinary: Facebook processes 930 million photo uploads, six billion 'likes' and 12 billion messages each day.[2] Google receives 40,000 searches every second, and YouTube users upload 100 hours of video to YouTube every minute.[3] According to IBM, 90 per cent of the data that exists today was generated in the previous two years.[4] The arrival of connected devices – the Internet of Things – and wearable technology is only intensifying this process, with important implications for sectors like Manufacturing and Health.

One consequence is the emergence of 'big data'.[5] This term was reportedly first used by technology consultants Gartner in 2001 to refer to datasets that are hard to manage using traditional database and analytical technologies.[6] The reasons for this are that these datasets have:

- **High volume**: They are too large to be stored in a single computer, or even on a single server. Instead, they need to be spread across an organisation's data infrastructure, using (sometimes multiple) computing clusters.

- **High variety**: Big datasets often draw on diverse sources and formats, including 'unstructured' ones like video, audio and text, which are very different from highly structured data, like financial information. Unstructured data has to be processed before it can be worked on, and is hard to store using traditional database architectures.

- **High velocity**: Big data is often generated at high frequency, and is most valuable when analysed and acted upon in 'real time'.

An ever-expanding array of technologies has emerged to help organisations deal with big data. They include Hadoop (a framework for distributed data processing), Cassandra (a big database system) and Hive (a big data warehouse), among many others.[7] These innovations, together with new developments in statistical and analytical methods such as machine learning, have made it easier to create value from data through better-informed decision-making and the creation of data-driven products and services. This is impacting on business performance in a way that we are only just starting to quantify.[8] In *Inside the Datavores* we showed how the 18 per cent of UK internet economy businesses who use their data more intensively are also, other things being equal, significantly more productive.[9]

THE DEMAND FOR DATA TALENT

Realising these new data opportunities requires the right skills: **engineering skills** to develop and maintain a reliable infrastructure to collect and process data, **analytical skills** to extract valuable knowledge from this data, and the **business know-how** to identify what questions to ask, and what to do about the answers.[10]

As more data becomes available and its value more apparent, these skills have become increasingly sought after. But is there enough data talent to go around?

Available evidence suggests the answer is no. According to McKinsey Global Institute, by 2018 the US will experience a shortfall of 140,000–190,000 'deep analytical experts' and 1.5 million data-savvy managers.[11] And recent research in the UK points to a 41 per cent increase in demand for 'big data professionals' in 2013 alone.[12] The same study suggests that 75 per cent of vacancies in big data occupations are hard to fill. The average salary for experienced big data professionals is almost double the UK average (and almost a third higher than the average for IT professionals).

NESTA'S RESEARCH PROGRAMME ON DATA ANALYSIS SKILLS

In this project we have partnered with Creative Skillset to study skills needs in one particular role, namely the data analysts charged with transforming data into business insight that drives action. The disruption this group is experiencing illustrates many of the wider workforce opportunities and challenges created by the data explosion.

Analysis of bigger and more varied data is driving innovation in all industries – be it customised search results and recommendation engines in the case of e-commerce, fraud detection in banking, robotics in manufacturing and transport, or genomics in health, to name but a few. However, harnessing many of these opportunities requires data analysts with a new combination of mathematics, statistics, computer programming skills and industry domain knowledge – they are increasingly referred to as 'data scientists'.[13]

Anecdotal evidence suggests that data scientists are exceptionally hard to find, and there are doubts about whether the UK's education system can rise to the challenge of producing what's needed.[14]

Our research seeks to answer three questions in order to inform policy and practice in this fast-moving field:

- What are the skills of productive data analysts?
- How easy is it to find those skills in the labour market?
- How can organisations manage their analytical talent to create the most value?

In July 2014, we published the first output of this research, *Model Workers*.[15] This report explored the questions above in 45 qualitative interviews with leading industry experts in the UK. It supported the idea of severe skills shortages in data analysis.

Although some of our interviewees were sceptical about terms like 'big data' (which some perceived as a fad) or 'data scientist' (which some saw as an opportunistic rebadging of previous professions), almost all were exploring the use of novel and 'messy' datasets in their businesses, and seeking – with only mixed success – talent with 'data scientist' profiles to do this.

Why were they struggling?

First, because available analysts lacked the right technical skills and industry experience. On the one hand, analysts straight out of education were not used to working with big data volumes and varieties: training them was costly and risky (not least because analysts would often leave to pursue other opportunities afterwards). On the other hand, experienced talent was expensive.

Second, data analysts lacked the right multi-disciplinary mix of skills: individuals who combined expertise in computing and analysis with knowledge of business were compared with unicorns, for all their rarity and mythical qualities.

The third reason for recruitment problems reflected considerations of internal business capacity: companies lacked the knowledge about how best to recruit analysts with advanced data analysis skills and, in some cases, senior management was not fully appreciative of the commercial value of data analysis skills.

ABOUT THIS REPORT

In the present report we set out to determine if the qualitative findings in Model Workers hold for a sample of 404 UK medium and large businesses in six sectors, and to outline what the results imply for government policy.

In addition to surveying these businesses, we extracted longitudinal financial data about them from Bureau Van Dijk's FAME database, and matched it with the survey responses to explore the link between data activities and productivity.

STRUCTURE

The structure of the report is as follows:

Section 2 describes the data we collected, and how.

Section 3 presents our findings. In sub-section 3.1, we describe the data activities and performance profiles of companies in our sample by segmenting them into four 'Data Groups'. In sub-section 3.2, we consider the labour market activities and outcomes for these four groups, and in particular the strategies they deploy to tackle workforce skills shortages and gaps. In sub-section 3.3, we analyse sector-specific differences in further detail.

Section 4 presents the conclusions of our analysis and their policy implications.

2. OUR DATA

SAMPLE DESIGN

Our sample was randomly drawn from the population of UK businesses with more than 50 employees in Bureau Van Dijk's FAME database.[16] We focused on medium- and large-sized business for three reasons:

1. Previous research suggests that the levels of analytics adoption among UK SMEs is very low (less than 1 per cent do any big data work).[17]
2. We were interested in learning about the organisation and management of data analysts. Some of these issues are less relevant for small organisations with simpler structures.
3. We wanted to match survey responses with financial data, but the availability of this data is patchy for companies with less than 50 employees (as they are exempt from reporting that information to Companies House).

We sampled companies from six sectors: Creative Media, Finance, ICT, Manufacturing, Pharmaceuticals and Retail, encompassing a range of domains where the literature indicates that data analysis is having – or could have – a substantial impact.[18]

QUESTIONNAIRE CONTENT

Our questionnaire is based on our previous Datavores research looking at skills and data practices, as well as the UK Commission for Employment and Skills Employer Skills survey.[19] It covers the following areas:

- **The data value chain**:[20] This includes questions on:
  - Data collection: what types of data are collected by companies, and how comprehensively.
  - Data processing and analysis: The methods that companies use to manage and analyse that data.
  - Data application: The areas where data analysis and insight create value inside the business.
- **Data talent sources**: What are the talent pools and disciplines that companies tap on for their analytical talent.
- **Labour market experiences**: Recent recruitment experience and the nature of hard-to-fill vacancies and skills shortages.
- **Skills strategies**, including investments in tools and training to upgrade workforce skills.
- **Data talent management practices**: The practices that respondents use to create value from their analytical talent, including team make-up, organisation, communication and compensation structures.

In order to ensure relevant responses, we implemented a screener at the beginning of the survey to exclude from the sample companies where data currently plays a negligible role, and where there is no intention of building up an analytical capability in the coming years. As a consequence, our sample is in no way representative of the general population of UK companies with 50 employees and above in the FAME database, nor is it our intention to conduct statistical inference about the UK business population.

The questionnaire was tested in in-depth cognitive interviews to ensure clarity of language and relevance of options. By the conclusion of the fieldwork, it took respondents just under 22 minutes to complete it.

FIELDWORK

The survey was carried out over the telephone by two research agencies – IFF Research and Ipsos MORI, between 5 March and 16 June 2014.

Data analysts may be found in any number of different departments within organisations so our interviewers made initial contact with Human Resource (HR) managers. This was based on the reasoning that HR managers were most likely to have information about the situation of data analysts inside the business. Where HR managers felt unable to answer the survey questions, the interviewers asked to be referred to someone else in the business.

Individuals were encouraged to participate in the research with a lottery draw for an iPad Mini.

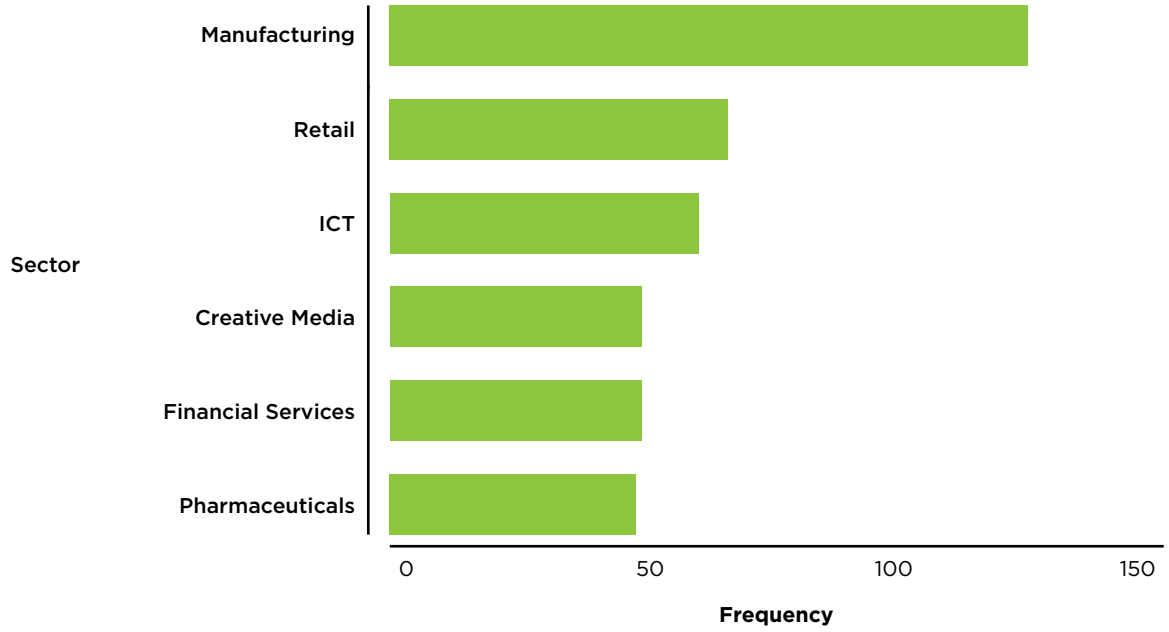
OUTCOMES

We collected data from 404 businesses. The response rate was 19 per cent and the eligibility rate (companies that, having accepted to participate, passed the screener), 73 per cent.

Around two-thirds of respondents work in HR functions (other functions represented in the sample include Finance, IT and operations). Almost two-thirds of respondents work in senior roles (such as Manager, Director, Managing Director and Chief Executive Officer).

Figure 1 shows the distribution of responses by industrial sector.

FIGURE 1: **SECTOR DISTRIBUTION OF SURVEY RESPONDENTS**



Manufacturing is over-represented, which is as expected given this sector's disproportionate presence in the population of medium- and large-sized firms. We obtain the lowest number of responses (49) in Pharmaceuticals, which is also expected, given the relatively small number of large companies in that sector.

FINANCIALS

As mentioned previously, we matched the survey responses with longitudinal financial data from FAME. This included a variety of indicators such as turnover, employment, costs of goods sold, average remuneration, and different measures of profitability. There were some inconsistencies in data availability caused by flexibility in the financial reporting timeframes required of UK businesses. For 49 per cent of companies, the last available accounts had been filed in 2013, and for 48 per cent, in 2012 (generally on the 31 December).

3. FINDINGS

3.1 DATA GROUPS: THEIR DATA VALUE CHAIN AND PERFORMANCE

WHY SEGMENT?

Instead of looking at the skills requirements of a 'typical company', we distinguish between the behaviours and performance of companies making very different uses of data. That there is no 'one size fits all' approach to data analysis and use was a strong lesson of our *Model Workers* report, and it is strongly echoed by our survey exercise too.

We segmented the companies in our sample based on their responses to questions about the 'data inputs' they worked with, and their attitude towards data (see Appendix 1 for details of the clustering analysis). We used three metrics for the segmentation:

- Data variety: How many data sources does a company use, and with what intensity?
- Data volume: How big (in volume) are the datasets that the company tends to work with?
- Data drive: To what extent does the company use data and analysis over experience and intuition to make business decisions?

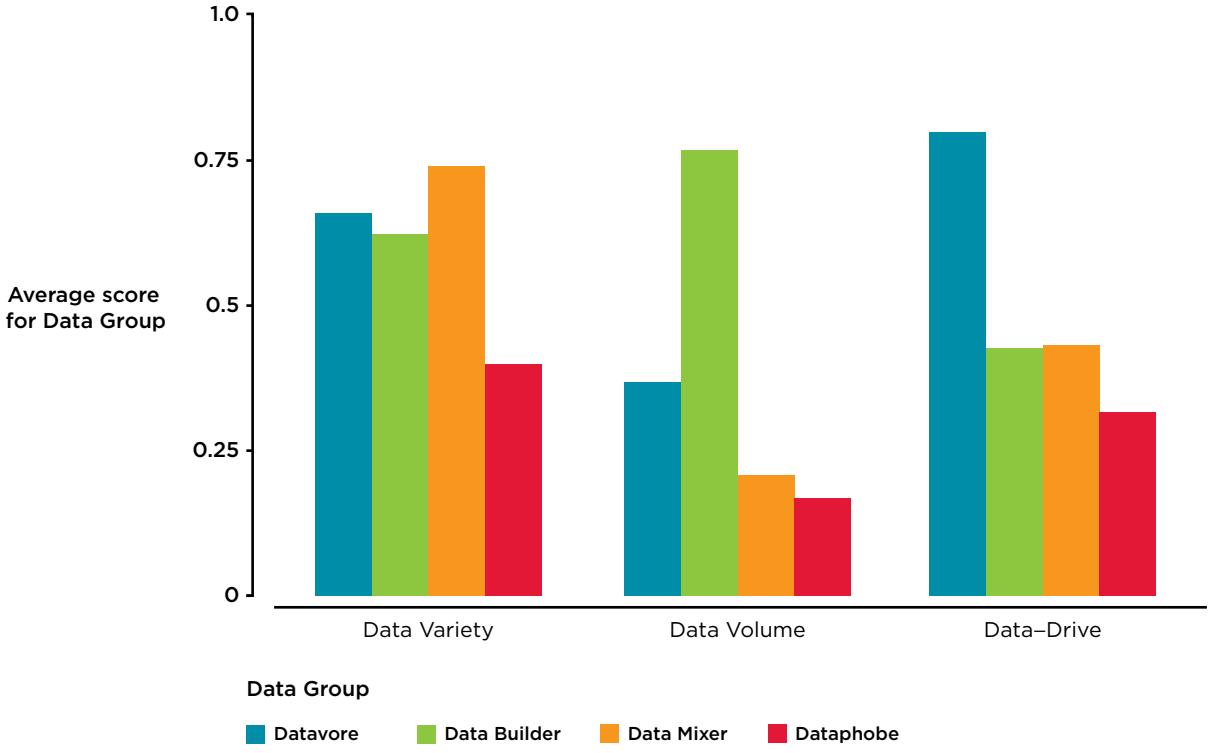
Our analysis revealed four distinct 'Data Groups' in the sample. They are as follows:

1. **Datavores (16 per cent of the sample)**: These are companies that make strong use of data and analysis for decision-making.[21]
2. **Data Builders (22 per cent of the sample)**: Companies using 'big' datasets requiring dedicated servers or (possibly multiple) clusters for parallel processing.
3. **Data Mixers (31 per cent of the sample)**: Companies that collect and combine data from a variety of sources.
4. **Dataphobes (30 per cent of the sample)**: Companies that work with small datasets and few data sources, and do not use data or analysis to make decisions.[22]

Figure 2 presents for each Data Group the normalised mean score for each of the three different data activity metrics.

It illustrates clearly the defining features of the different Data Groups. But it also shows where there are overlaps in data practices: Datavores and Data Builders also combine a variety of data sources, for example; Datavores sometimes work with big data volumes, and Data Builders and Data Mixers also use data to make decisions.

FIGURE 2 **DATA GROUP SCORES IN DATA VARIETY, VOLUME AND DRIVE**



Companies working with few data sources, small data sets and who do not use data or analysis for decision–making – 'Dataphobes' – account for a considerable number of companies in our sample. In the remainder of the report, we use the term 'Data–active' to characterise all other firms excepting these Dataphobes.

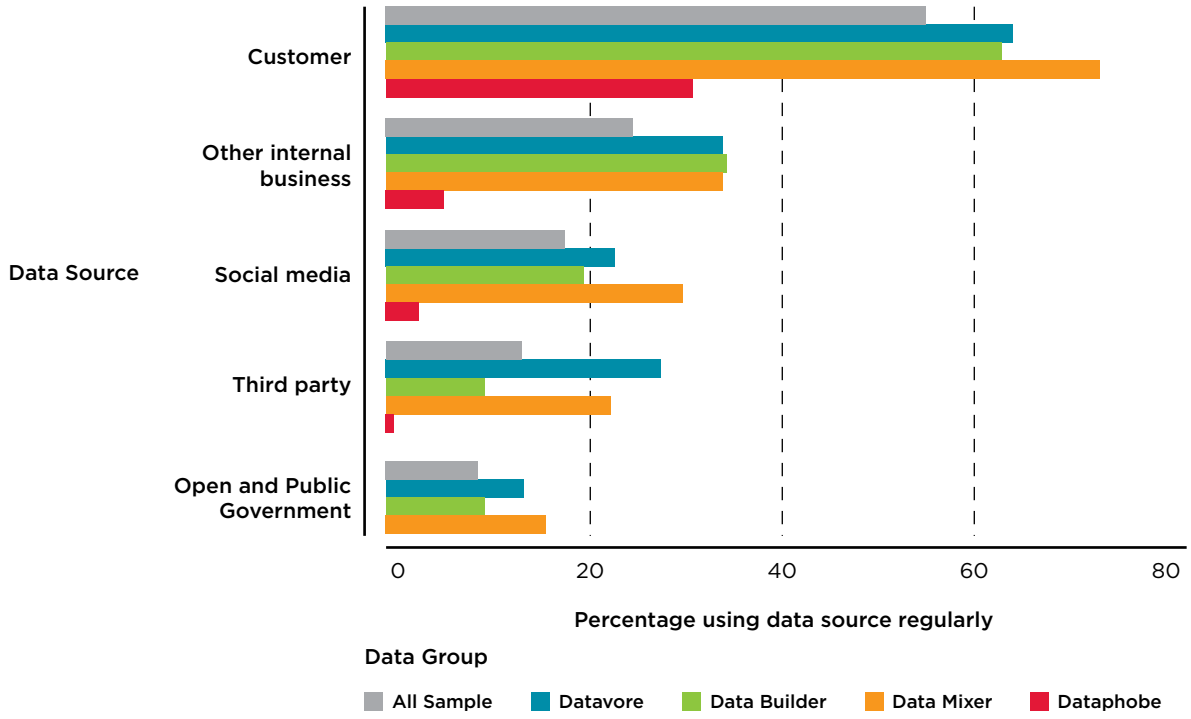
HOW DO THE DATA VALUE CHAINS OF OUR FOUR DATA GROUPS COMPARE?

In order to gain a better understanding of the differences between companies in these Data Groups, we explore their behaviour at different stages of the data value chain: specifically, can we detect differences in the data sources they are using, in their approaches to analysis, and in the nature of the impacts of data analysis on their business?

Data collection: Figure 3 shows for each Data Group, and the sample as a whole the extent to which companies regularly use different sources of data. Customer data is by far most regularly used by all companies, followed by other internal business data (e.g. financial and operations data), social media data, third party data (e.g. from credit scoring agencies, audience research etc.) and last, but not least, open and public government data (which is used routinely by 9 per cent of respondents).

Naturally, Data Mixers make more regular use of a variety of data sources compared with other firms. Dataphobes, on the other hand, rarely make use of data sources other than on their customers.

FIGURE 3 **REGULAR USE OF DIFFERENT DATA SOURCES**



Data analysis: Figure 4 shows what tools different companies use to manage their data, and what analytical techniques they employ to extract insights from it (see the Glossary for their description).

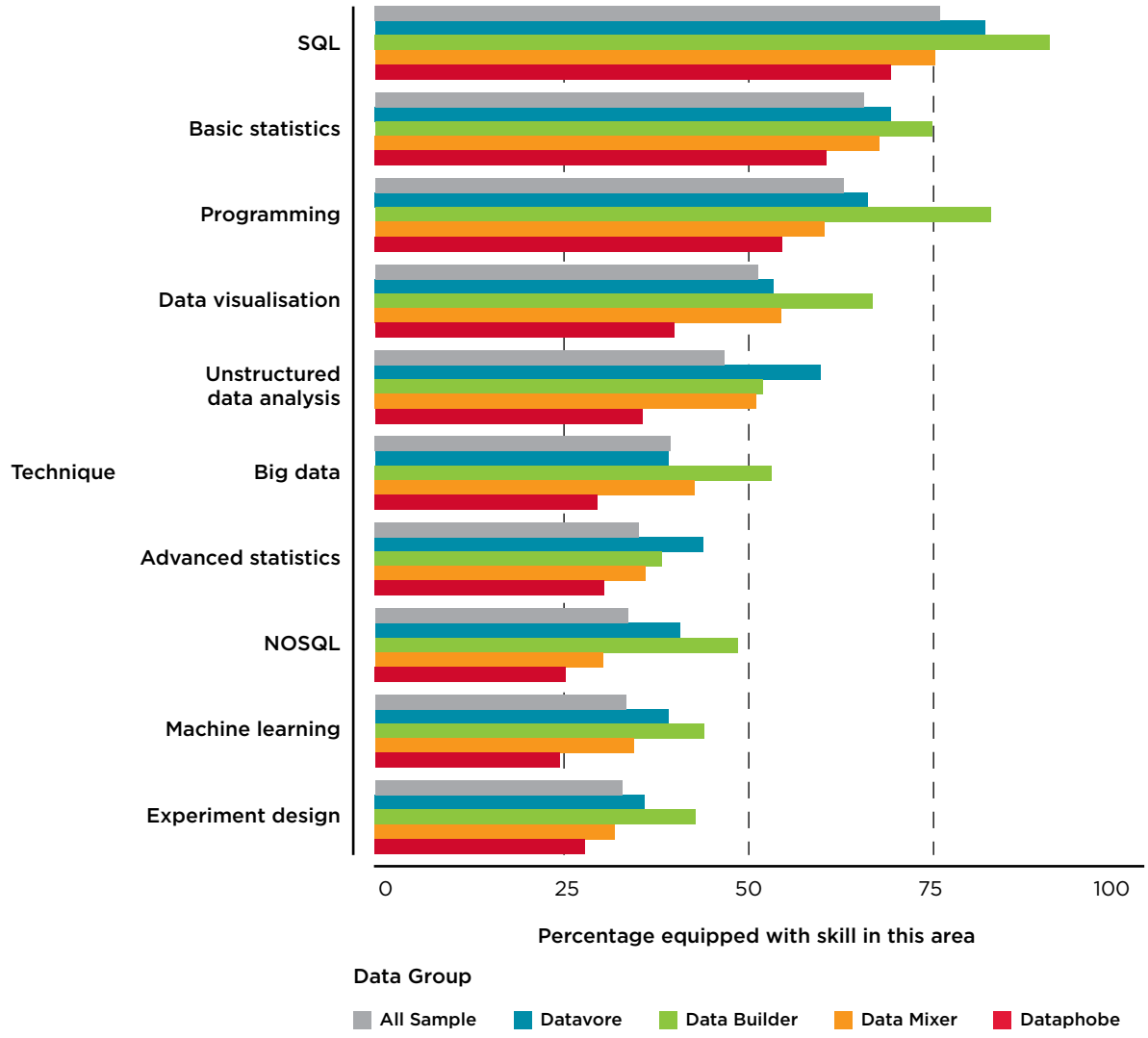
Unsurprisingly, older technologies and methods are more prevalent than newer tools: Over half of all respondents have SQL, basic statistics and general programming capabilities. Half report that they can do data visualisation.

A substantial minority of companies (around one–third) have skills in more technically complex areas like advanced statistics and machine learning, however. In general, Data Builders are more likely to make use of data processing and analysis tools. They are 30 per cent more likely than the average to have the skills to work with big data (such as skills with the Hadoop parallel data processing framework), and machine learning skills, and 45 per cent more likely to have NoSQL skills to store and manage unstructured data. All of this makes sense, given that they work with high volume datasets requiring a strong IT infrastructure and automated data processing.

We also find a considerable number of Datavores and Data Mixers with advanced statistics and unstructured data analysis skills.

As expected, Dataphobes lag behind in the adoption of all data management and analysis techniques.

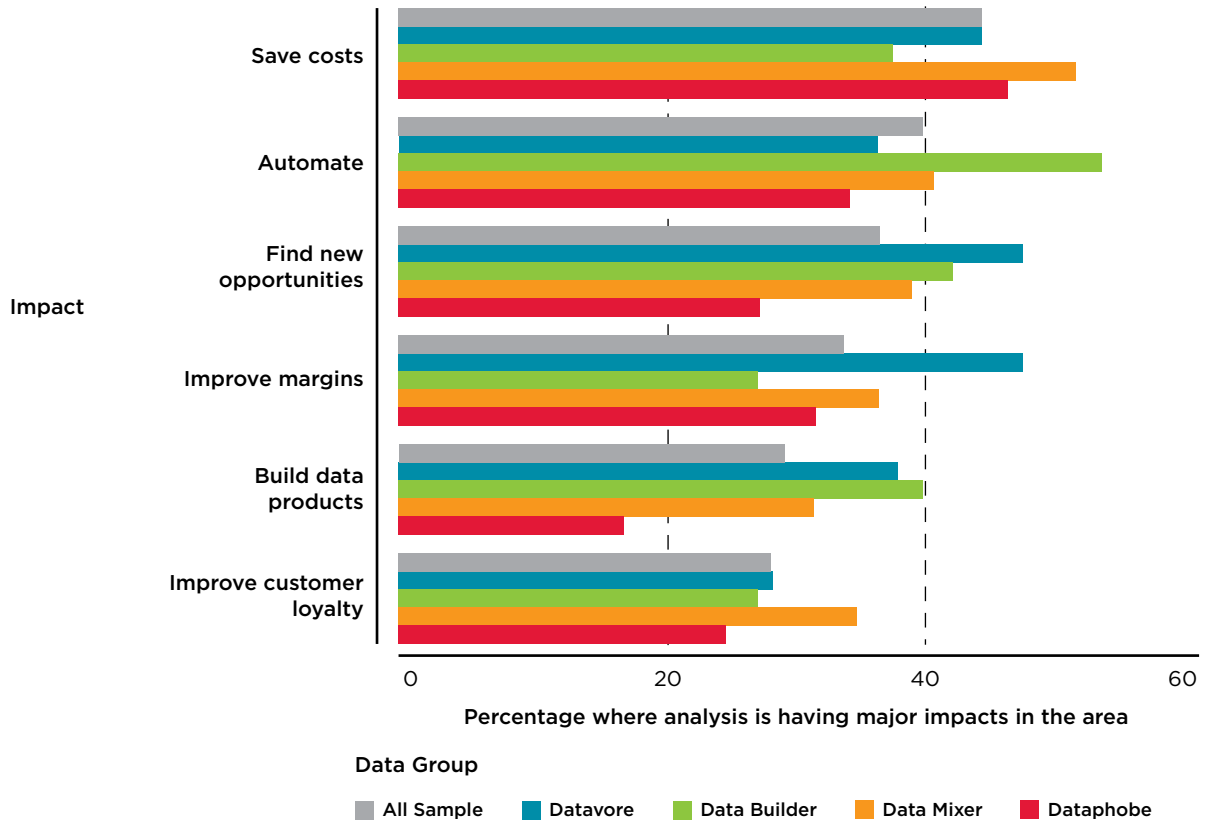
FIGURE 4 **SKILLS IN DATA MANAGEMENT AND ANALYSIS TECHNIQUES**



Data application: Figure 5 shows the business areas where, according to our respondents, the work of data analysts is having a major impact. Although process-related improvements (cost saving and automation) are mentioned most often, firms report impacts in many other parts of their business too.

It also shows an interesting split between Dataphobes who are primarily using analytical insights to save costs, automate and improve margins, and Data-active companies, where analysis contributes to the discovery of new opportunities (this is particularly the case for Datavores), creates revenues from data products (especially for Data Builders) and improves customer loyalty (as reported by Data Mixers). This pattern is consistent with the idea that firms that are less advanced with data use it incrementally, with a focus on efficiency, whereas companies further along their data journey make it part of their innovation and strategic processes.[23]

FIGURE 5 **AREAS OF IMPACT FOR DATA ANALYSIS**



DATA-ACTIVE COMPANIES ARE FOUND IN ALL SECTORS

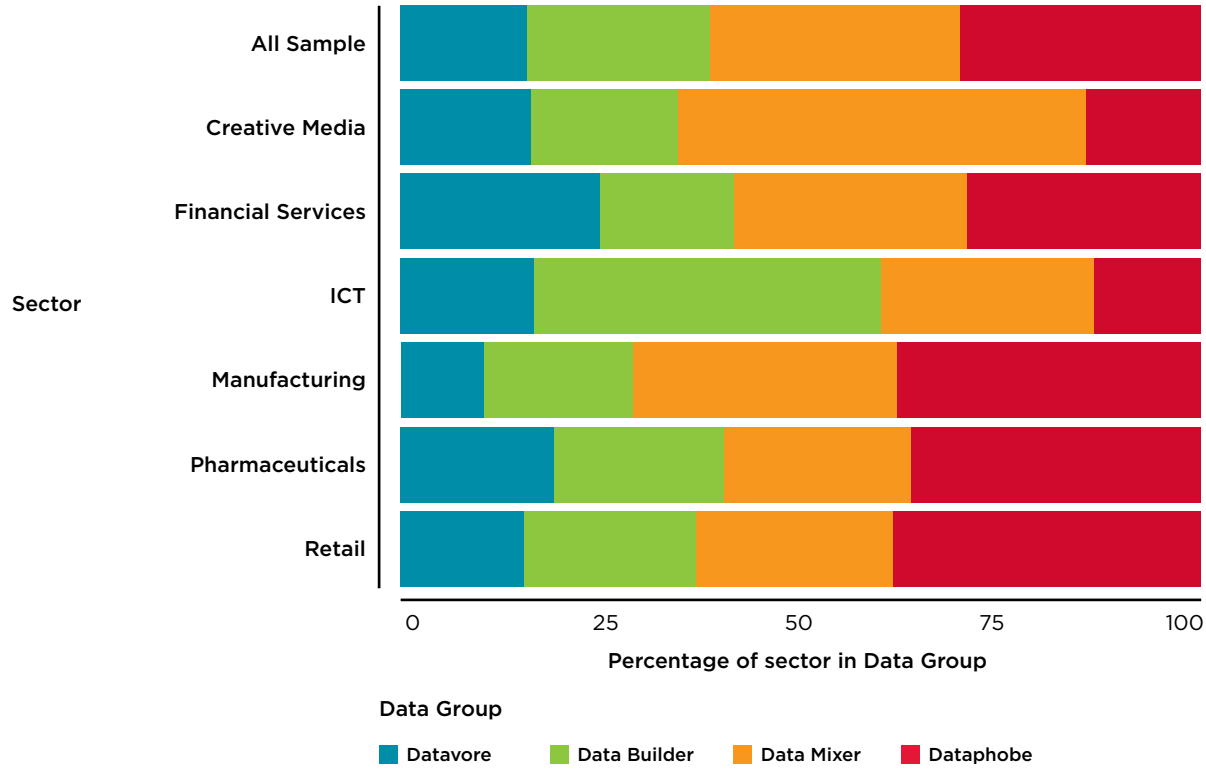
It is generally acknowledged that the benefits of data are not confined to a single industry. On the contrary, data is transforming all sectors, reflecting the 'general purpose' nature of digital technologies.[24]

Figure 6, where we look at the distribution of our Data Groups across industrial sectors confirms this view, with Data-active companies in all sectors.

That being said, some of the Data Groups are more prevalent in some sectors. So, it turns out that there are more Datavores in quant-heavy Financial Services and science-intensive Pharmaceutical companies, for example. Data Builders, working with big datasets, are more commonly found in ICT. Intuitively, there are more Data Mixers in the Creative Media sector, where a variety of web sources (social media, audience data, web analytics) are being combined to create value.

We find Dataphobes in all sectors too, although interestingly they are somewhat more common in sectors that on the whole have – at least up until now – been less disrupted by the internet (Manufacturing, Retail and Pharmaceuticals).[25] We might speculate that Dataphobe businesses in these sectors will see particular challenges in coming years, with the growth of technologies that are connected and data-enabled, like the Internet of Things, e-Commerce and wearables.

FIGURE 6 **DATA GROUPS BY SECTOR**



DATA-ACTIVE COMPANIES ARE MORE INNOVATIVE IN THEIR PRODUCTS AND EFFICIENT IN THEIR PROCESSES

What is the quantitative link between a company's data activities and its performance? Answering this question can help us gauge the contribution of data analysis skills to business value, and the obstacles to growth created by bottlenecks in the supply of analytical talent in the UK.

Figure 5 presented firms' perceived impacts of analysis on their business. In Figure 7, we consider whether a company sees itself as a product innovator 'in general' (i.e. if it tends to launch products and services ahead of its competitors), and whether it considers itself efficient in its business processes.

The survey responses suggest that Data-active companies perform much better than Dataphobes on both counts.[26]

In particular, Data-active companies are more likely to self-identify as being product/service innovators.[27] Data Builders – the companies most likely to describe themselves as product/service innovators – are over 50 per cent more likely than Dataphobes to say they launch products and services ahead of their competitors.

Figure 8 considers whether respondents think they are more efficient than their competitors. Once again, Data-active companies report higher levels of process efficiency than Dataphobes, though the differences are less marked.[28] Datavores in particular strongly agree that they are more efficient in their processes (almost 44 per cent do, compared with only a quarter of Dataphobes).

FIGURE 7 **REPORTED LEVELS OF PRODUCT/SERVICE INNOVATION BY DATA GROUP**

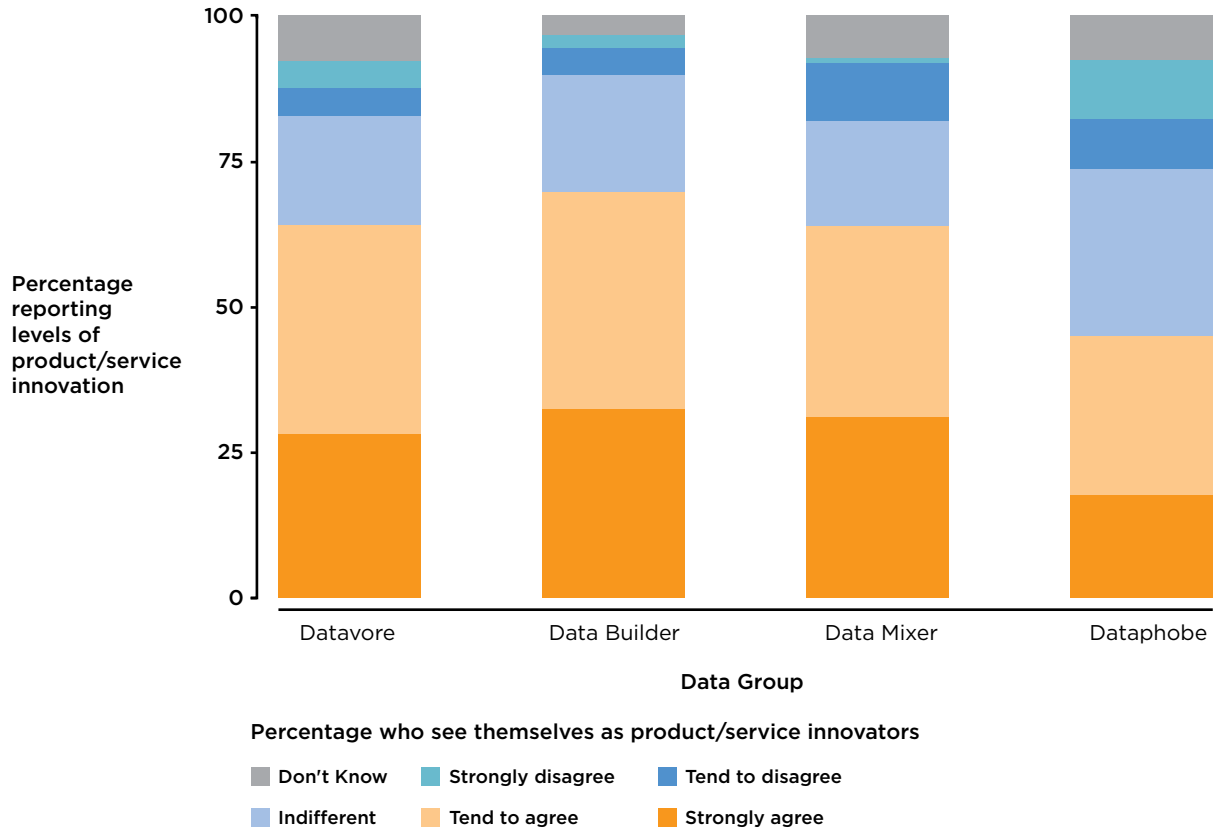
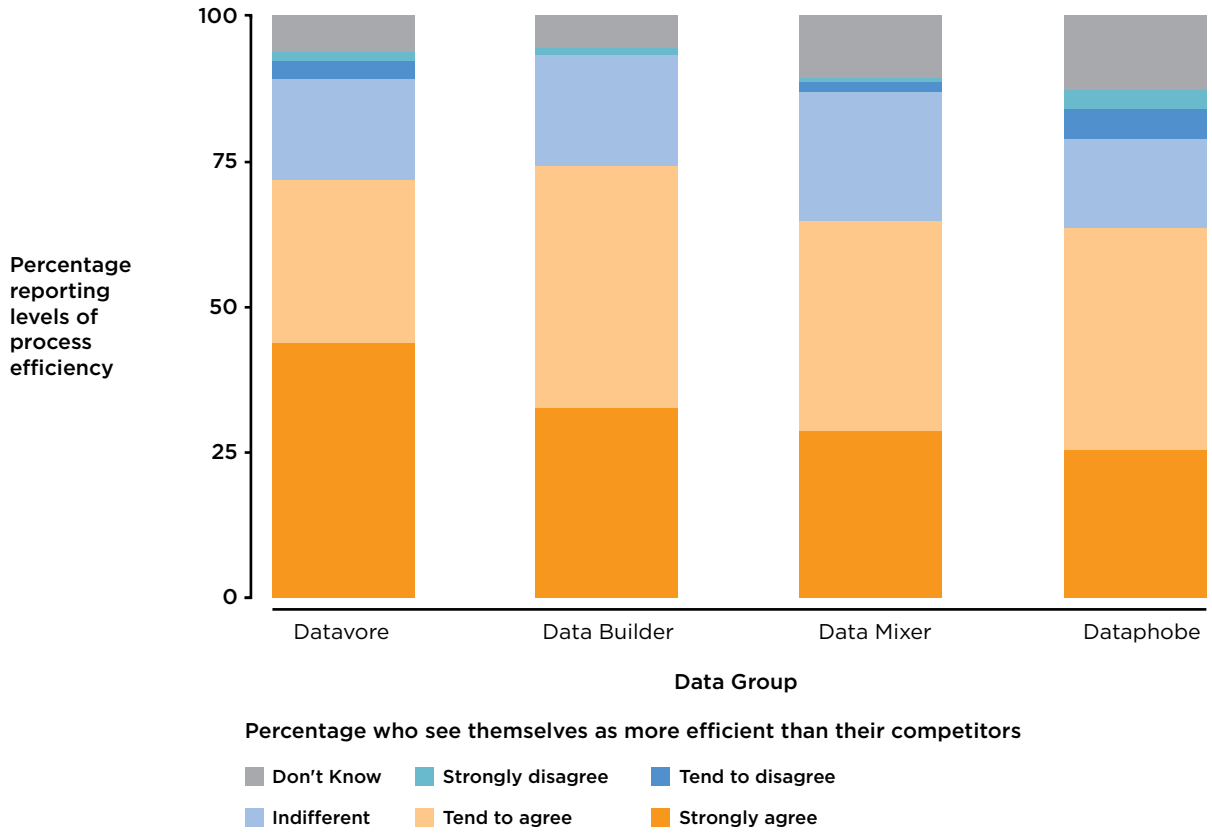


FIGURE 8 **REPORTED LEVELS OF PROCESS EFFICIENCY BY DATA GROUP**



DATA–ACTIVE COMPANIES ARE ALSO MORE PRODUCTIVE

Self–reported measures of performance like those we have just described obviously suffer from subjectivity. Furthermore, simple bivariate correlations (in the case above, between Data Group status and company performance) fail to consider the possibility that a third variable might be driving the relationship that has been observed.

To address these issues, we carried out an econometric analysis of the link between a company's Data Group and its economic performance using financial data from FAME, following the same modelling strategy in Inside the Datavores (see Appendix 2 for further information). Our goal was to determine if Data–active companies were, other things being equal, more productive than Dataphobes.

Figure 9 shows the results of this analysis for three models where we progressively introduce more variables to control for other determinants of productivity (measured by value added).[29]

1. The **Factor Controls** model only controls for a company's levels of production input, namely employment and capital.

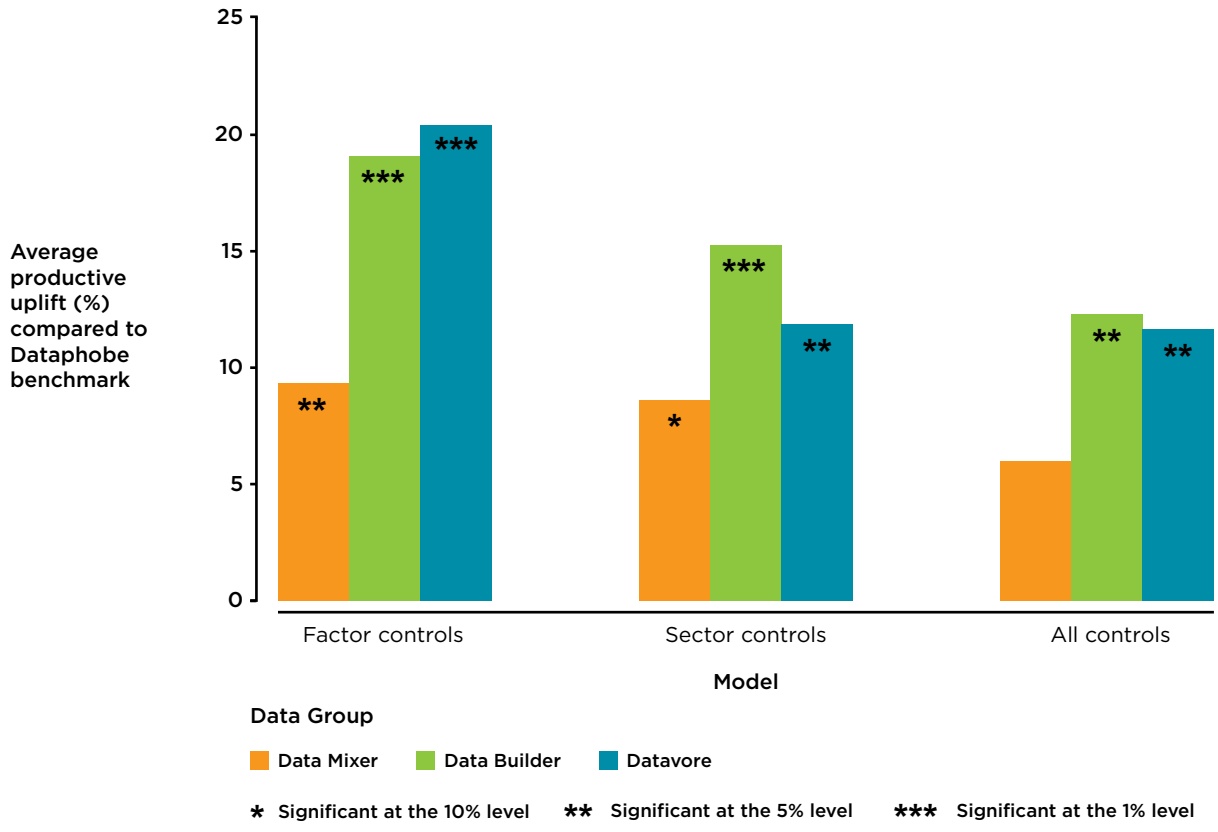
2. The **Sector controls** model adds controls for the industry a company operates in.

3. The **All controls** model adds controls for company age and innovativeness, based on the survey responses to the aforementioned questions about product innovation and process efficiency. The idea is to account for some 'unobserved' differences between companies that might correlate with both productivity and data activity (like investment in Research and Development, or managerial bent towards risk–taking), thereby inducing a false impression of causality.

The height of the bars show the average differences between companies in a given Data Group and the Dataphobes 'benchmark', while the stars show the statistical significance of the estimate (three stars are significant at the 1 per cent level, two stars are significant at the 5 per cent level and one star is significant at the 10 per cent level). (see Appendix 2 for the full regression results).

Our estimates show that companies in all three Data–active groups – and especially Datavores and Data Builders – are significantly more productive than the Dataphobes, even after we control for potentially confounding variables. As might be expected, progressively adding more controls diminishes the marginal contribution of data to productivity, however, and in the case of Data Mixers makes them statistically insignificant (albeit still positive).

FIGURE 9: PRODUCTIVITY ACROSS DATA GROUPS



The headline result is that, other things being equal, companies in the Datavores and Data Builders group are, on average, over 10 per cent more productive than Dataphobes. One plausible interpretation of this finding, consistent with the previous discussion, is that Datavores and Data Builders use analysis to identify new commercial opportunities, improve their efficiency and generate revenues from innovative data products and services, and that this drives up their value added.[30]

3.2 ACCESS TO ANALYTICAL TALENT AND SKILLS

The preceding analysis shows that higher levels of data activity are in broad terms associated with stronger company performance in terms of higher rates of product innovation, levels of process efficiency, and productivity. The implication is that analytical skills shortages will create bottlenecks for these activities, for superior company performance, and growth.

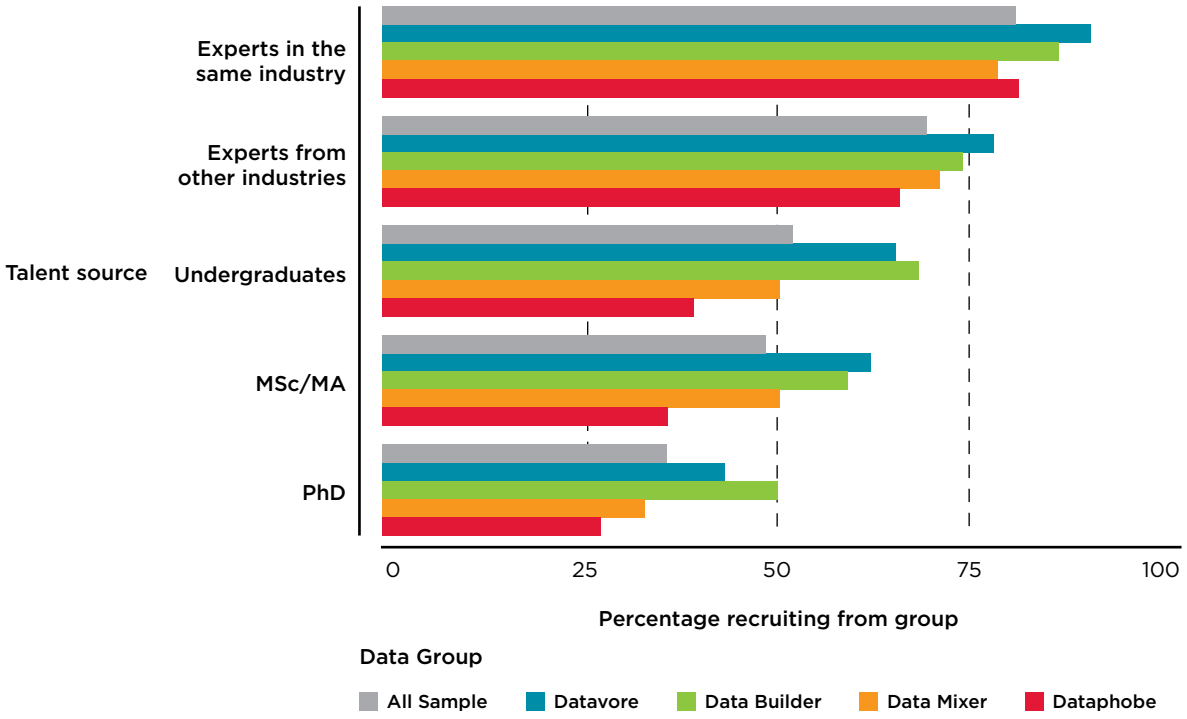
In this sub–section, we provide evidence on the labour market for data talent. In particular we use our survey findings to look at the talent pools that companies in different Data Groups draw on, the disciplines they bring together to build an analytical capability, and their recent recruitment activity and difficulties filling vacancies. We also identify the specific skills areas for which shortages and gaps exist, and the types of training which UK businesses are investing in to keep the skills of their data analysts up to date.

ALTHOUGH MOST COMPANIES PREFER TO RECRUIT EXPERIENCED TALENT, THEY HIRE FROM UNIVERSITIES TOO.

When asked about the types of analytical recruits they take, most companies (8 in 10) mention ‘experienced talent from inside their own industry’, followed by ‘experienced talent from other industries’ (mentioned by almost 7 in 10) (Figure 10). A considerable number of businesses (36–50 per cent) are also hiring analysts straight of university, including undergraduates, postgraduates and PhDs (over a third of respondents mention doctoral programmes as a source of talent).

Datavores and Data Builders recruit from universities more often. Around 75 per cent in the first group and 70 per cent in the second identify at least one academic source of talent, compared with less than one half of Dataphobes. This is consistent with the importance of advanced research skills in much analytical work, as well as the challenges of finding more experienced talent (which we return to later on).[31]

FIGURE 10 **SOURCES OF ANALYTICAL TALENT (PER CENT RECRUITING FROM SOURCE)**



DATA-ACTIVE COMPANIES ARE BUILDING UP A 'DATA SCIENCE' CAPABILITY

Figure 11 shows the academic disciplines that our respondents recruit from. Business disciplines (including Management, Marketing and Finance) are the main sources of analytical talent (almost two-thirds of companies hire analysts with this background).

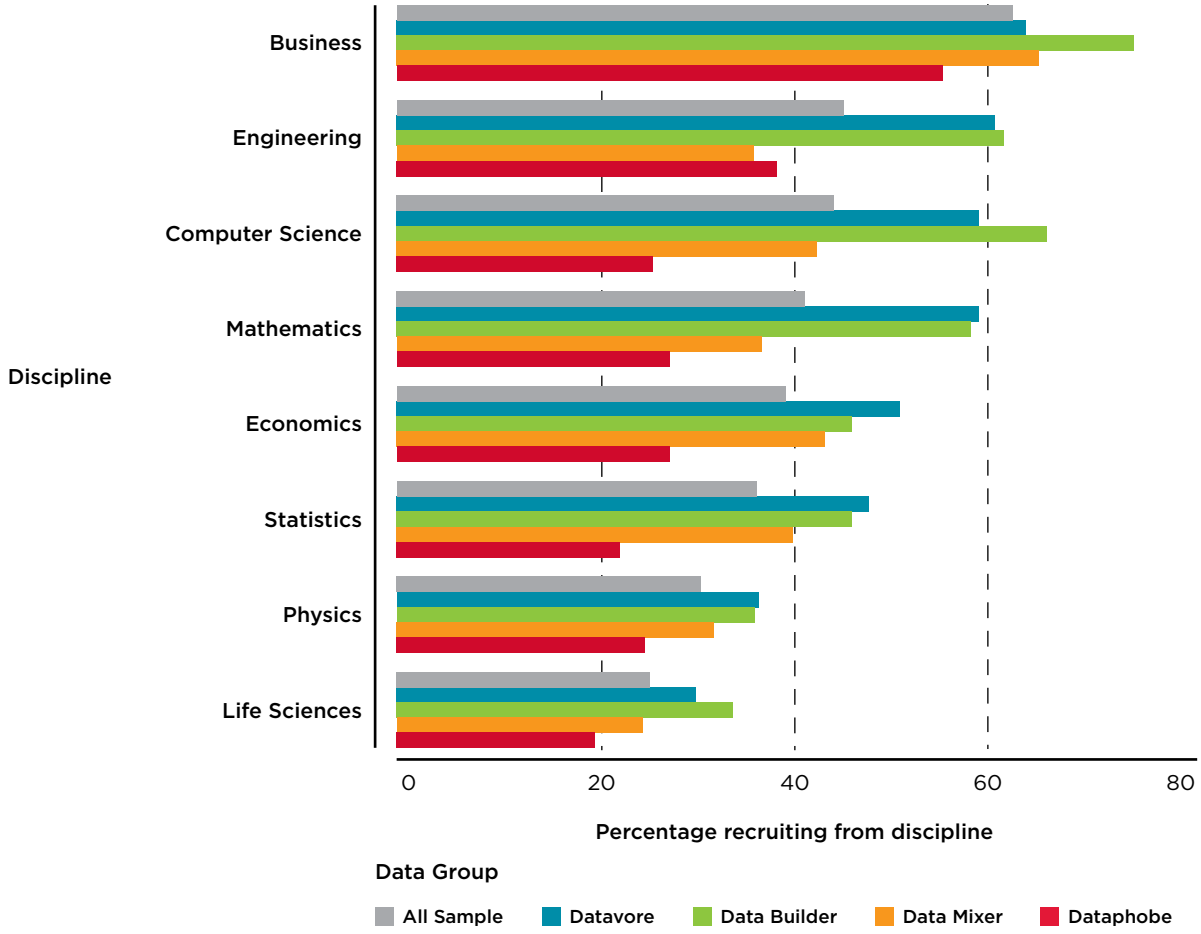
Datavores and Data Builders are much more likely to recruit candidates with a Computer Science, Engineering or Mathematics background – around 6 in 10 do, which is double the figure for Dataphobes. Between 40 per cent and 50 per cent of the companies in the Data-active Group more generally (which also includes Data Mixers) hire statisticians.

This mix of computer programming and quantitative knowledge makes a good deal of sense for these companies: they work with more complex and harder to manage datasets, use more sophisticated analytical techniques, and deploy data to build new products and services. This requires developing a 'data science capability' that brings together coding and analysis. Around one-half of Datavores and Data Builders, and a third of Data Mixers, hire analysts from core numerical disciplines (Mathematics or Statistics) and Computer Science at the same time, compared with just 17 per cent of Dataphobes.

In *Model Workers* we quoted industry leaders stressing the importance of multidisciplinary analytical teams for innovation and learning. Our survey findings strongly echo this, with almost everyone in the sample (86 per cent) agreeing that 'team diversity is important for

performance'. We find that Data-active companies stand by this rhetoric in recruiting from a wide range of disciplines. The median Datavore and the median Data Builder hire from four disciplines simultaneously – twice as many as the median Dataphobe.

FIGURE 11 **THE DISCIPLINES OF DATA ANALYSTS**



DATA-ACTIVE COMPANIES ARE HIGHLY ENGAGED IN THE LABOUR MARKET FOR ANALYTICAL TALENT...AND EXPERIENCE GREAT DIFFICULTIES RECRUITING

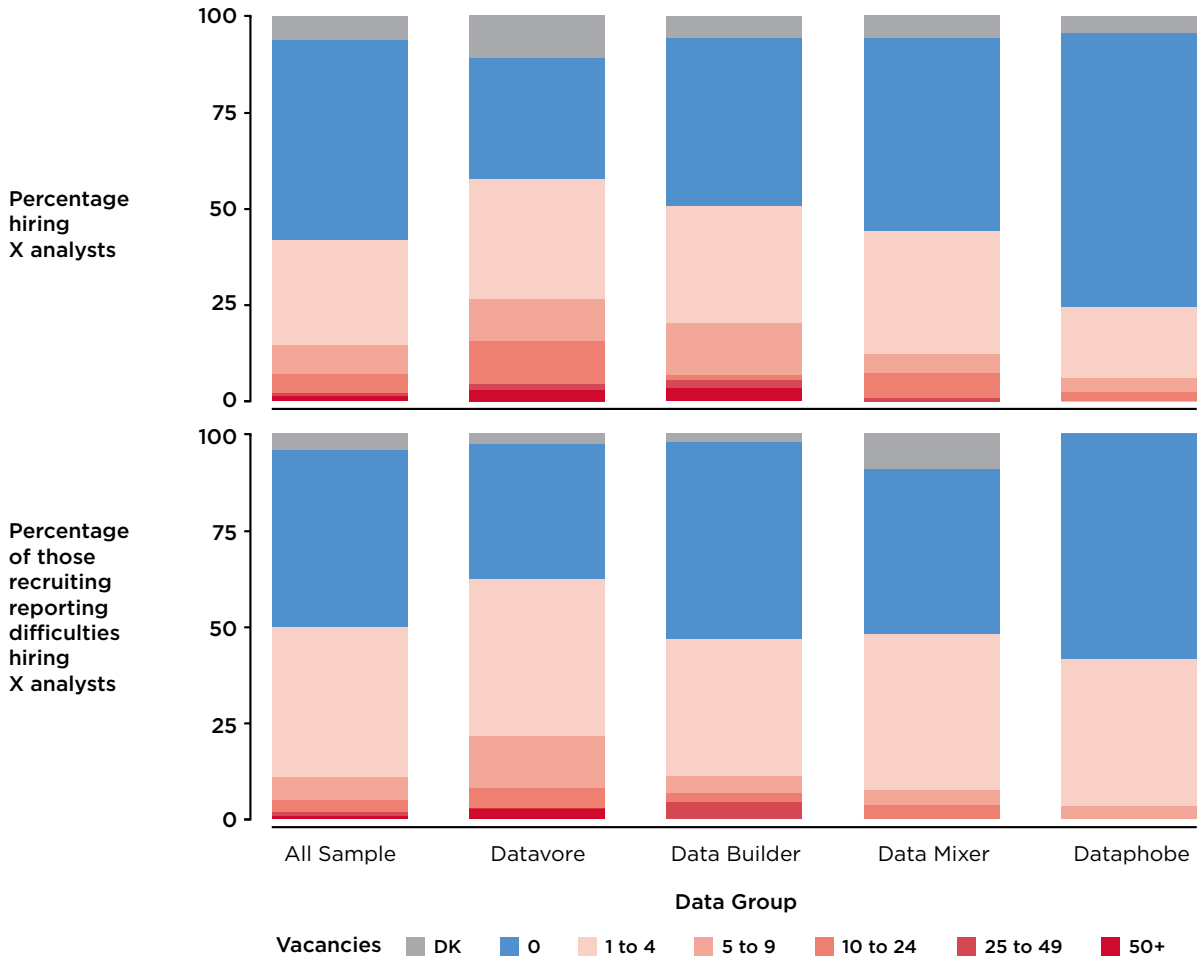
Figure 12 shows recent levels of analyst recruitment in our sample, and the difficulties they face in filling vacancies. The top panel displays the percentage of companies that sought to fill at least one analyst vacancy in the 12 months prior to being surveyed (we use different shades of red to illustrate the number of vacancies they sought to fill), while the bottom panel displays the percentage of companies that suffered recruitment difficulties (again, the shades of red show the number of vacancies that were hard to fill).

The headlines are:

1. **Data-active companies are much more engaged in the market for analytical talent.** While just over 40 per cent of the companies in the sample have sought to hire a data analyst in the 12 months before the survey, the equivalent percentage for Datavores is 59 per cent, and for Data Builders 52 per cent. Almost one-half of Data Mixers have sought to recruit data analysts in the previous 12 months. These Data-active companies are also recruiting greater numbers – almost 20 per cent of Datavores have sought to recruit ten data analysts or more in the previous year, compared with just 2.3 per cent of Dataphobes.

2. **Around one-half of the companies that have sought to recruit data analysts report difficulties filling at least one vacancy.** Here, Datavores are particularly likely to highlight difficulties (almost two-thirds of those who sought to recruit have experienced problems in filling at least one vacancy.) Fewer Dataphobes claim to suffer in this area: only 4 in 10. These differences across groups are also visible in the volumes of hard-to-fill vacancies involved: around 15 per cent of Datavores and 10 per cent of Data Builders struggle to fill five vacancies or more, compared with 3.4 per cent of Dataphobes. These findings are worrying because they suggest that many high-performing Data-active UK businesses are struggling to access the analytical talent they need to exploit their data. Finding analytical talent seems to be less of an issue for Dataphobes which, as we showed in previous sections, are less sophisticated in the analysis of their data, and innovative in its applications.

FIGURE 12 **LEVELS OF RECRUITMENT AND HARD-TO-FILL VACANCIES**

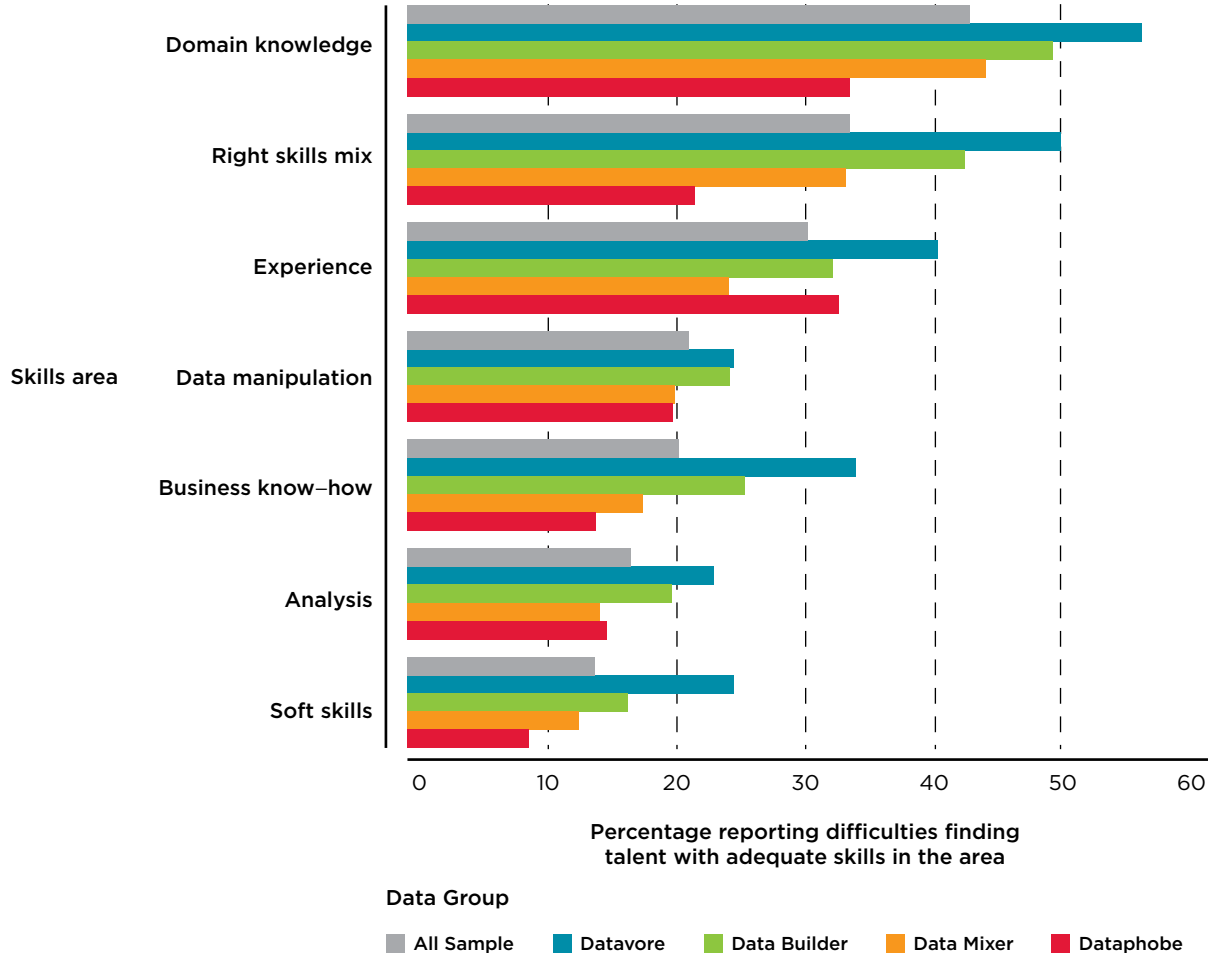


ANALYSTS WITH ADEQUATE DOMAIN KNOWLEDGE, THE RIGHT SKILLS MIX AND EXPERIENCE – THE 'PERFECT ANALYST'– ARE HARDEST TO FIND

Figure 13 shows the proportions of respondents reporting difficulties in finding analysts with the right level of skills and knowledge in different areas.

It shows that the main areas of concern are insufficient domain, or industry, knowledge, talent lacking the right mix of skills, and professional experience – between 30 per cent and 40 per cent of all companies flagged up issues in these areas. A smaller (although still substantial) percentage of companies – between 15 per cent and 20 per cent – mentioned difficulties finding people with the right technical skills (that is, data analysis and data manipulation).

FIGURE 13 DIFFICULTIES FINDING DATA TALENT WITH THE RIGHT SKILLS IN DIFFERENT AREAS

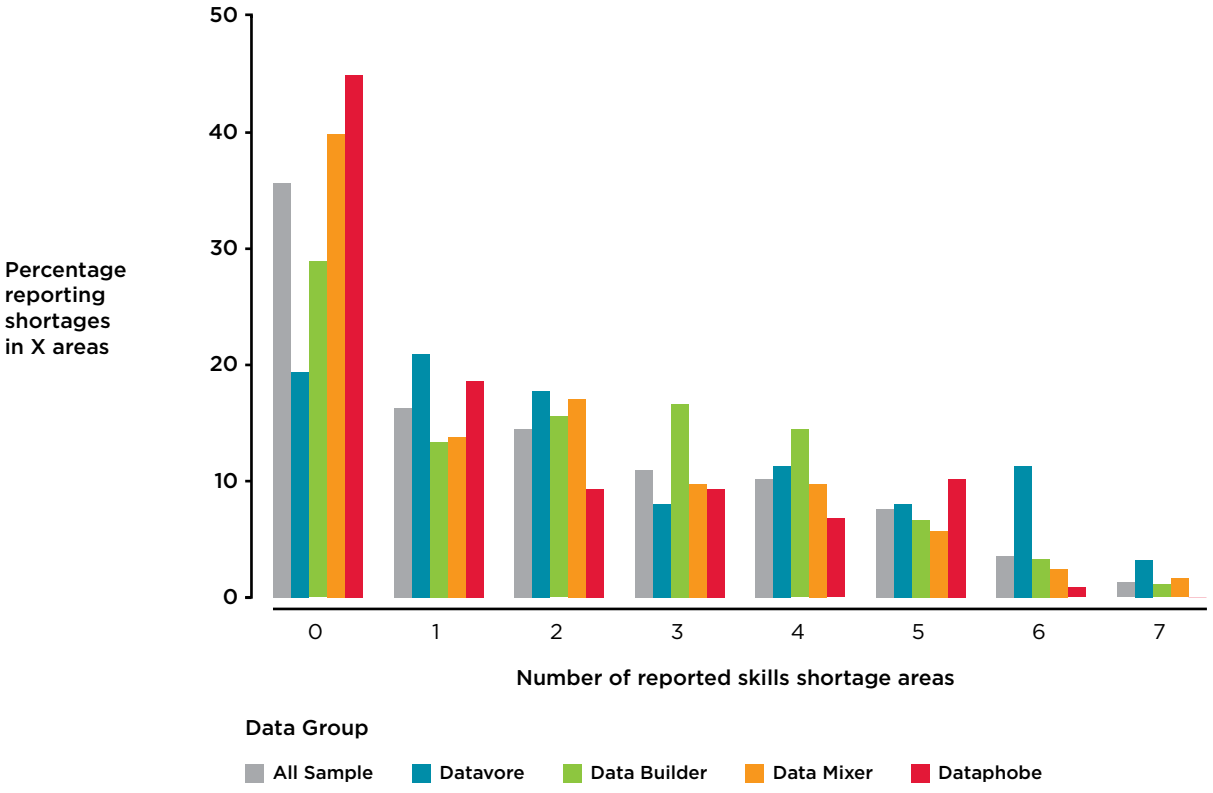


When comparing the responses of different Data Groups, it is striking that Datavores and Data Builders are more likely to report difficulties finding data talent with the right skills in all skills areas. As Figure 14 shows, 80 per cent of Datavores and around 70 per cent of Data Builders report difficulties in at least one skills area (compared with 56 per cent of Dataphobes.) One–third of Datavores and one–quarter of Dataphobes identify four or more skills shortage areas.

Datavores and Data Builders are particularly concerned with finding data talent with relevant domain knowledge (over half report problems in this area), and with the right mix of skills (flagged up by 50 per cent of Datavores, and over 40 per cent of Data Builders). Insufficient business know–how is a proportionately bigger issue for Datavores and Data Builders than other companies too.

These findings strongly reinforce a conclusion in *Model Workers*: namely that data analysis is becoming increasingly embedded in a wide range of industries and business processes, but in order to create value from it, it needs to be combined with domain knowledge and business know–how – a 'data science' mix that, our findings confirm, is in short supply. These constraints are particularly biting for Datavores and Data Builders who are doing more complex analysis and working with larger volumes of data and who, the results of this report suggest, are likely to be forgoing value added because of the recruitment difficulties they face.[32]

FIGURE 14 **NUMBER OF SKILLS SHORTAGE AREAS BY DATA GROUP**



MOST COMPANIES SEE OPPORTUNITIES TO UPGRADE THE SKILLS OF THEIR WORKFORCE

Figure 15 shows our respondents' views about the current skills levels of their workforce in different areas, and the perceived opportunities and needs for further development.

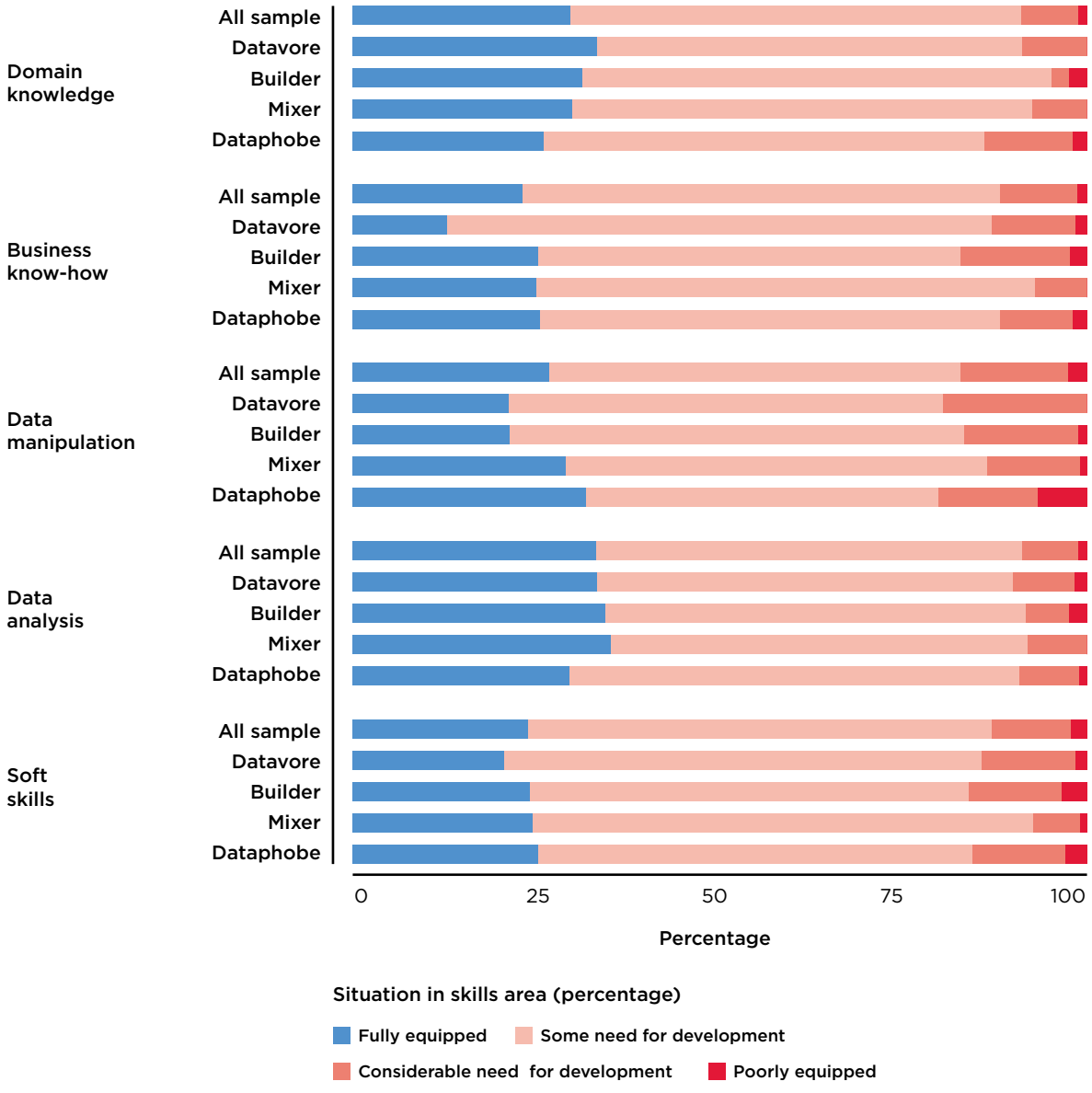
In general terms, the companies we surveyed see extensive opportunities to develop the skills of their analysts. Few think their workforce is adequately equipped to address the

needs of their business today. The areas where our respondents are least satisfied are data manipulation skills, business know-how and soft skills (roughly in that order).

Intuitively, unlike the case of recruitment, respondents are relatively less concerned about the domain knowledge of their workforce – expected, given that current employees pick up some of this knowledge on the job.

Interestingly, we do not find major differences between Data-active companies and Dataphobes in their reported satisfaction with workforce skills, suggesting that the low levels of Data Activity among the Dataphobes do not reflect insufficient skills in their analytical workforce; rather, they are the result of a lack of analytical ambition.

FIGURE 15 **DEVELOPMENT NEEDS IN DIFFERENT SKILLS AREAS**

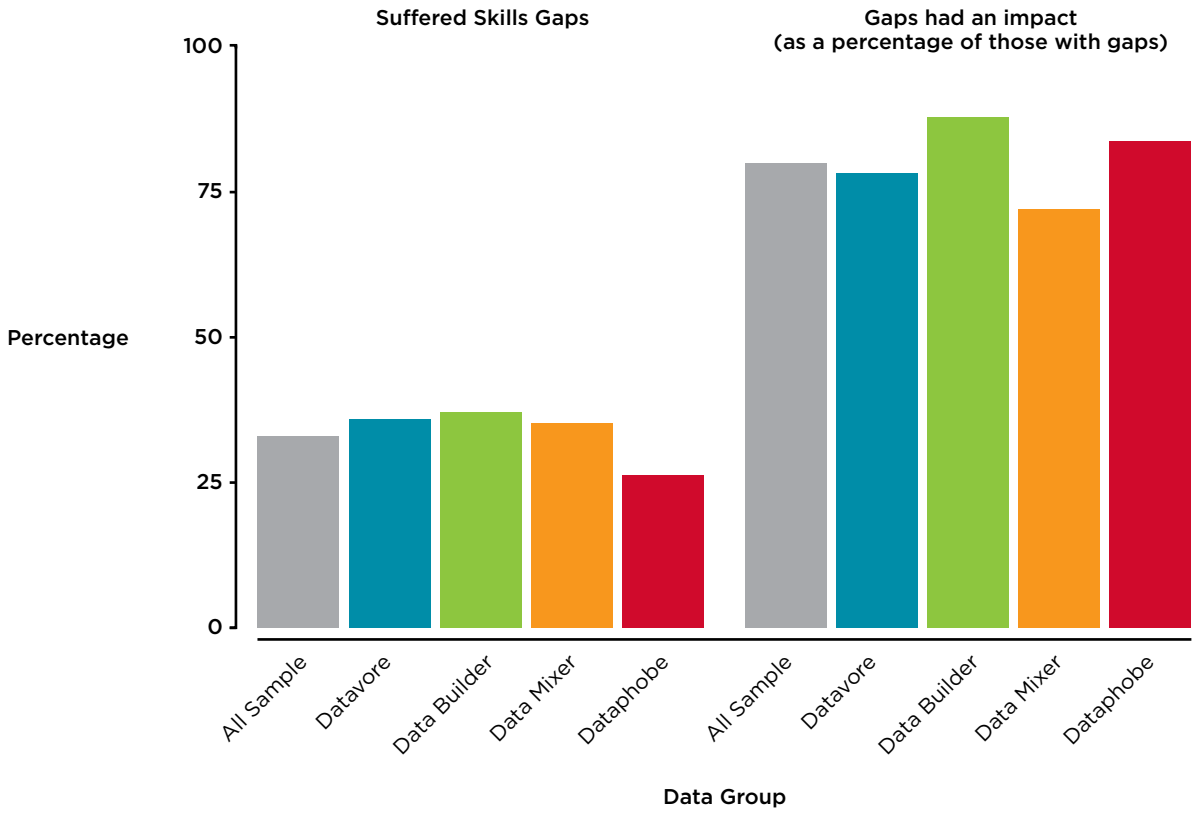


ONE-THIRD OF COMPANIES REPORT SKILLS GAPS IN THEIR ANALYST WORKFORCE, AND OVER THREE-QUARTERS OF THESE SAY THESE GAPS DETRACT FROM THEIR PERFORMANCE

Leaving aside specific skills areas where there are opportunities for development, Data-active companies which are more ambitious in their uses of data are more aware of skills gaps in their current workforce. They are between 40 per cent and 50 per cent more likely than Dataphobes to report such gaps (35 per cent–37 per cent say this is the case, compared with 26 per cent of Dataphobes, see left-hand panel in Figure 16).

The right-hand panel in Figure 16 shows the perceived impacts of these gaps in the companies suffering them: over 70 per cent of all companies and almost 90 per cent of Data Builders say that skills gaps in their workforces are having a negative impact on company performance.

FIGURE 16 **SKILLS GAPS AND THEIR IMPACTS**



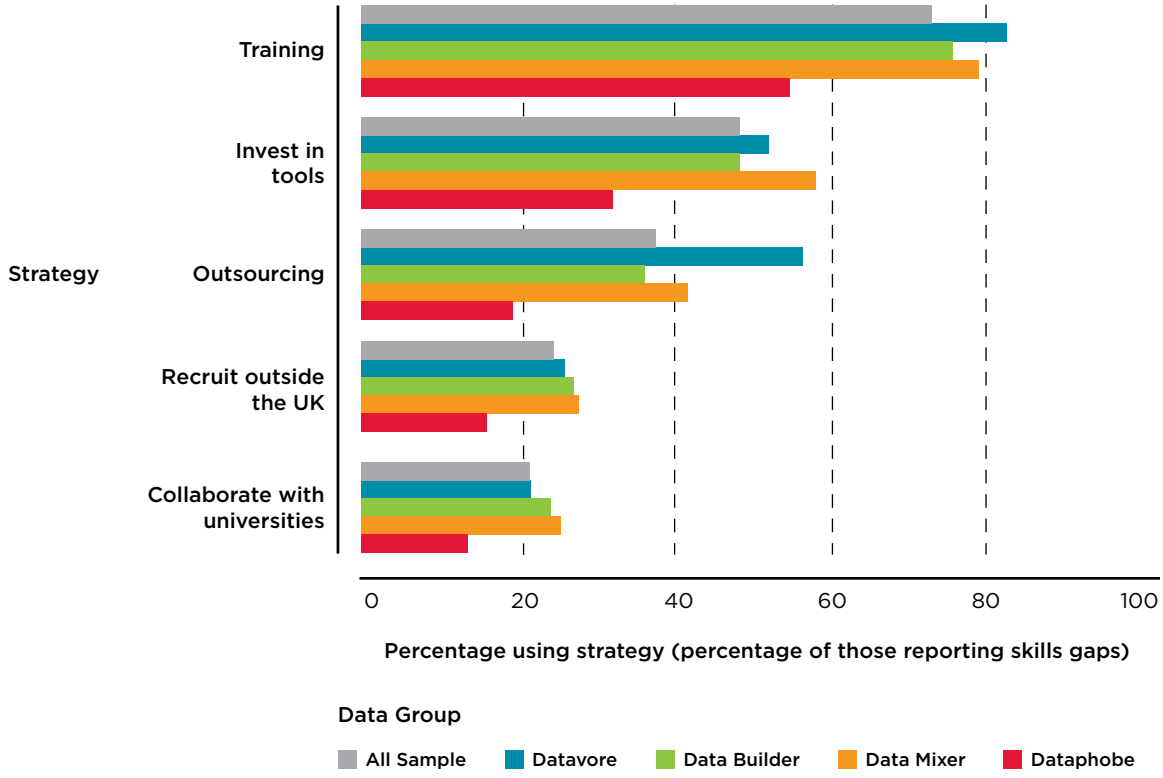
COMPANIES IN OUR SAMPLE ARE USING MULTIPLE STRATEGIES TO ADDRESS SKILLS GAPS

The companies we interviewed in *Model Workers* described a number of strategies they were adopting to deal with analytical skills gaps. These included investing in software tools that substituted for data analysts (or made them more productive), outsourcing analytical work to specialist providers, employing talent from overseas, and working with universities to spot talent and train it.

Figure 17 shows the frequency with which different strategies are used by survey respondents (note that the base for these percentages are only those companies who reported skills gaps, not everyone). The main action being taken is to increase training budgets, and expand trainee programmes, underscoring the importance that firms attach to developing their in-house analytical capabilities – 73 per cent are doing this. In comparison, fewer companies are resorting to external analytical skills via outsourcing (the exception here are Datavores – 56 per cent of those with skills gaps are looking for specialist skills outside the company).[33]

It is also worth noting that a substantial minority – just over 20 per cent – are working with universities to address their skills gaps and around one-quarter are seeking to address them by tapping into overseas talent.

FIGURE 17 **STRATEGIES TO ADDRESS SKILLS GAPS IN THE WORKFORCE**



COMPANIES REPORT HIGH LEVELS OF TRAINING, OFTEN INVOLVING INNOVATIVE PEER-TO-PEER AND ONLINE FORMATS. UNIVERSITIES ARE THE LEAST POPULAR SUPPLIER OF TRAINING

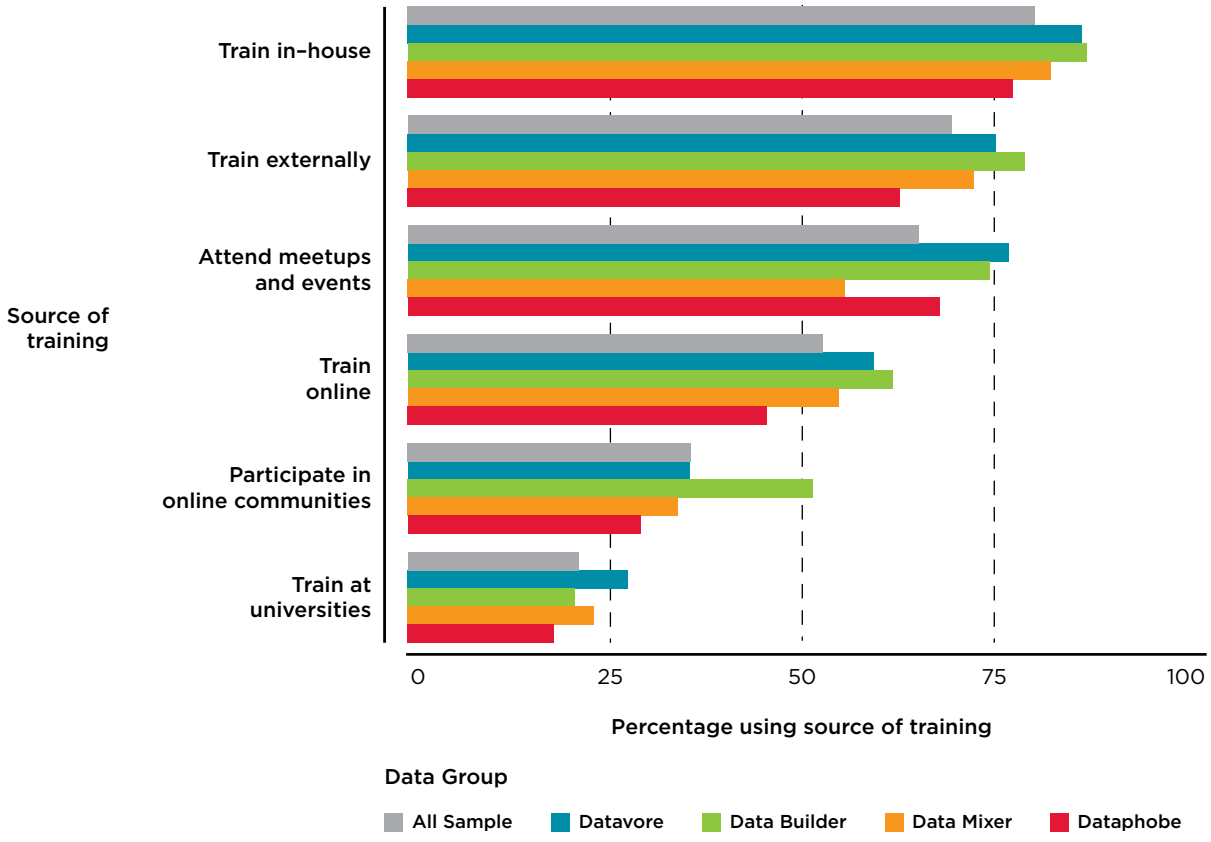
The data landscape is changing rapidly, and this requires employers to make substantial investments in workforce skills if they are to stay abreast of the most recent developments in the field. As Figure 15 showed, most companies see opportunities to build up the skills of their analytical workforce in a number of areas.

Almost 80 per cent of our respondents reported providing in-house training and 70 per cent making use of external training (Figure 18).[34]

Going beyond this, more innovative, web- and peer-based training sites are proving remarkably popular (almost as popular as external training). Two-thirds of companies say their analysts participate in events and meet-ups, more than half use online training services and Massive Open Online Courses (MOOCs), and one-third contribute to online communities and open source software projects, including competition sites such as Kaggle. Notably, universities are the least popular source of training – with only just over one in five companies mentioning them.

These findings are consistent with the fact that the data science field is one where rapid change in experimental tools and data sources – often freely available online – have resulted in a vibrant culture of informal, peer-to-peer learning informed by 'hacker' traditions.[35]

FIGURE 18 **TRAINING ACTIVITIES AND SOURCES**



3.3 SECTORAL PROFILES

To conclude our findings section, we have compared data activities and analytical skills issues in different sectors. We summarise the outcomes of this exercise in Table 1 (focusing on the Data Value Chain) and Table 2 (focusing on access to analytical talent and skills). (See Appendix 3 for the figures underpinning this comparison.)

SECTORAL VARIATIONS IN THE DATA VALUE CHAIN AND SOURCES OF TALENT REFLECT DIFFERENCES IN MARKET STRUCTURES, BUSINESS MODELS AND TYPES OF DATA USED

There are good reasons for expecting variations in data analysis practices across sectors, depending on market structures and business models. For example, social media data should be expected to be more important for consumer-facing Creative Media companies using platforms like Twitter or Facebook for promotion and to measure audience behaviour, while Pharmaceutical companies are more likely to be reliant on open and public data (including data published by public research organisations). And this indeed is what we find (Table 1).

A sector's choice of data tools also depends on the types and volumes of data it works with. For example, ICT companies processing large data sets will have a need for non-relational database technologies like NoSQL and big data technologies like Hadoop. In terms of impacts, consumers in Creative Media markets value novelty, so it should not be surprising that analysts in that sector generate impacts by using data to find new opportunities. By comparison, B2B manufacturers prioritise analysis to identify cost efficiencies.

The pattern of differences is similar when we look at the analytical talent piece: for example, the survey findings suggest that science-based Pharmaceutical companies are more likely to recruit from PhD programmes. And in terms of disciplines, more ICT companies recruit computer scientists, while Financial Services firms value economists, and Manufacturers seek out data-expert engineers.

Regarding perceived skills shortages, we see that Creative Media companies are particularly concerned with insufficient domain expertise in the talent pool. This might plausibly be explained by the fact that, until recently, the sector has not been considered as a natural career destination for quantitative analysts, which means that knowledge of the sector in the analytical labour market could be expected to be low. Heightened concerns in the ICT sector about a lack of experienced analysts, and analysts with the right skills mix probably reflects the fast rates of change in the big data technology landscape (which make it difficult to find workers proficient with the latest data platforms and tools), as well as a dearth of multi-skilled data scientists.

We also see intuitive differences across sectors in attitudes to training, with ICT companies that operate in a fast moving technology landscape investing more in data training, including peer-to-peer and web-based solutions and communities that may be the only source of information about cutting-edge technologies and tools. Pharmaceutical companies that have stronger links with academic research departments, it turns out, use universities for training more often.

CREATIVE MEDIA, ICT AND FINANCIAL SERVICES HAVE BEEN MORE ACTIVELY RECRUITING DATA ANALYSTS

In terms of their labour market activity, the Creative Media, ICT and Financial Services sectors have been most active in their recruitment of data analysts. Almost two–thirds of Creative Media companies, and more than one–half of ICT and Financial Services companies had tried to hire at least one analyst in the previous 12 months. Manufacturing and Pharmaceuticals, arguably the least 'web–active' sectors in our study, recruited less often, with just over one–quarter of all companies trying to hire at least one analyst recently. Pharmaceutical companies were more likely than other sectors to report difficulties filling at least one vacancy: Around seven in ten did, compared with just under half in other sectors.

DESPITE THESE SECTORAL DIFFERENCES, WE SEE REMARKABLE LEVELS OF DATA INNOVATION AND TALENT 'CROSSOVER' IN THE SAMPLE

The literature abounds with examples of surprising and innovative transfers of data sources and analytical methods across domains –like the application of models developed by physicists to finance, epidemiology to interbank settlements, or machine learning to translation.[36]

We see hints of this sectoral and interdisciplinary crossover in the data: for example, 10 per cent of Manufacturing companies report using social media data regularly, and almost 30 per cent of retailers claim to have skills in experimental design (i.e. setting up randomised control trials to test the effects of different strategies – e.g. in marketing – on website users). Four in ten retailers say that analysis creates an impact through the development of data–based products and services, and a third of Creative Media companies use it for automation (as we are seeing in markets like advertising, via real–time bidding for ad inventory).

The strength of talent and disciplinary flows across industries are also striking: seven in ten companies in our sample are recruiting experienced analysts from outside the 'natural talent pool' in their industry. Over one–third of Creative Media companies hire Physicists, for example, and one–third of Pharmaceutical companies recruit Economists. A quarter of Retailers employ analysts from Life Science backgrounds. These behaviours underscore the transferability of quantitative skills across domains, and the opportunities for serendipitous 'combinatorial' innovations created by unexpected talent flows across sectors.

TABLE 1 **SECTORAL DIFFERENCES IN THE DATA VALUE CHAIN**

Area	Observations
Data Variety	Creative Media companies use the widest variety of data sources, followed by ICT and Financial Services.
Data Volume	ICT Companies work with the biggest datasets, followed by Financial Services and Creative Media.
Data Drive	There are no big differences across sectors in this variable, although Financial Services are, on average, the most data-driven in their decision-making, and Manufacturing the least.
Data collection	Creative Media companies tend to use a wider variety of data sources regularly (and in particular social media data), while ICT businesses work with more third party data. Pharmaceutical companies are the biggest users of open and public data.
Data management and analysis	Creative Media companies have database skills, and work with unstructured techniques like social network analysis and text mining more often. ICT companies are more adept with general programming languages, as well as big data, machine learning and NOSQL technologies. Financial companies are bigger users of advanced statistical techniques, followed by Pharmaceuticals. Pharmaceutical, ICT companies and Retailers make more use of data visualisation tools.
Data impacts	Creative Media businesses use data to identify new opportunities, and improve margins and customer loyalty, while manufacturers and retailers apply it to reduce costs and improve efficiency. ICT are (with Financial Services) the main data-driven 'automators' and, together with Creative Media and Retail, the sector where data products are having the biggest impact.

TABLE 2 **SECTORAL DIFFERENCES IN THE SKILLS AREA**

Area	Observations
Analytical talent sources	Financial Services show the biggest propensity to recruit analysts within their own industry, while Creative Media and ICT look relatively further afield. Creative Media are the biggest recruiters of undergraduates and masters graduates, and Pharmaceuticals hire most frequently from PhD programmes.
Disciplines	Business disciplines are the most popular source of analytical talent in all sectors except Pharmaceutical (which prefers to recruit Computer Scientists). ICT are the leading recruiters of Computer Scientists and (with Manufacturing) Engineers, while Financial Services hire more Mathematicians, Economists and Statisticians. Pharmaceuticals hire more Physicists and people from Life Science backgrounds.
Labour market activity and outcomes	Creative Media, Financial Services and ICT are most active in the analytical labour market: more than half had sought to recruit in the last 12 months, while Pharmaceuticals and Manufacturing were least active. There were no substantial differences across sectors in numbers of hard-to-fill vacancies (just under half of all companies had difficulties), excepting Pharmaceuticals, where filling vacancies was harder (over 7 in 10 companies that sought to recruit report at least one hard-to-fill vacancy).
Areas with skills shortages	Lack of domain knowledge in the analytical talent pool was a particular concern for Creative Media companies (half of them flag this up), while ICT are struggling to find talent with the right skills mix, and experience. Data manipulation issues seem to be particularly severe for Creative Media, ICT and Pharmaceutical companies. Manufacturers were in general least concerned about skills shortages in the analytical talent pool.
Skills development	Retailers, Manufacturers and Pharmaceutical companies see more opportunities to develop the domain knowledge of their analysts, while ICT and Manufacturing are more concerned about their business skills. Manufacturers and Retailers would like to develop their data manipulation skills, and Pharmaceuticals their analysis skills.
Skills gaps	There are no big differences in skills gaps issues across sectors. ICT and Manufacturing companies are more likely to say their skills gaps are having an impact on their performance.
Sources of training	ICT companies are most active in the training space. This includes in-house, external and online training, as well as participation in online communities. Creative Media also make strong use of online training and participation in web communities, while retailers attend meet-ups and events. A larger share of Pharmaceuticals companies and Manufacturers use universities as a source of training.

4. POLICY IMPLICATIONS

Data is transforming the economy, increasing efficiency and creating new opportunities for innovation. We are constantly generating data, from changes in how we shop, communicate and meet, to the clothes we wear and the gadgets we use, and businesses and government are becoming more adept at creating value from this.

The UK, referred to by some as 'The Connected Kingdom' is particularly well–placed to benefit.[37] But if data is the new oil, logically, it won't be useful to business until refined. That requires analytical skills.

That much has been made clear in recent years by an extensive body of independent research by Nesta into the experience of the 'datavores' – those businesses that make heavy use of data for driving their business decisions – as well as the work of academics and other public bodies like the Tech Partnership.[38] In this report, we have shown a strong link between data, business innovation and productivity: Data–active companies are over 10 per cent more productive than 'dataphobes' that don't exploit their data, controlling for other determinants of productivity.

However, the data–active companies we surveyed are struggling to find suitable talent. Two thirds of the Datavores who tried to recruit analysts in the last 12 months struggled to fill at least one vacancy. A recent employer survey by the Tech Partnership shows that big data analytics is the tech occupation with the biggest skills gaps.[39] It seems that while data is part of the answer to the worrying productivity gap between the UK and other countries, barriers to accessing analytical talent might be preventing UK businesses from fully harnessing its potential.

By and large, the problem is finding people with the right mix of skills: the data scientists who combine technical skills, analytical and industry knowledge, and the business sense and soft skills to turn data into value for employers are very hard to find – so much so that some people refer to them as 'unicorns.'[40]

In the absence of such unicorns, businesses are building their analytical capability through multidisciplinary teams. Members of a team may have a number of core skills in common, and individuals will have specialist skills developed within particular disciplines. This underscores the need not just for multidisciplinary working, but for data analysts with strong teamwork and communication skills.

In *Analytic Britain*, a policy briefing that we have developed together with Universities UK, we set out policy actions to address this situation and ensure that UK businesses have access to the analytical talent they need to thrive in the data revolution, spanning the whole talent pipeline, including schools, colleges, universities and the labour market and industry. The recommendations aim to remedy skills shortages in the short term, while ensuring a sustainable supply of excellent analytical talent in the long term. Additionally, the

recommendations encourage cross-sector collaboration so that knowledge about how to create value from data and awareness of analytical skills shortages are not trapped in silos, but widely shared.

The data revolution has implications not only for experts with advanced analytical skills (i.e. data scientists), but for the entire workforce. We all need to become more data literate to operate successfully in increasingly 'data-rich' environments. This is a key lesson from *Count Us In*, a review of the landscape for quantitative skills in the UK published by the British Academy.[41] Our recommendations reflect the diversity of analytical skill levels which are needed, and also suggest creating early 'touch points' between young people and data, acknowledging that in some cases these will be the beginning of a life-long analytical career, while for others it will be part of feeling familiar and confident with data, whatever their final career destination.

The all-pervasive reach of the data revolution explains why a variety of disciplines and skills need to come together if the UK is to fully benefit from it. As a system challenge, it can only be addressed with a systemic programme of actions like the one we set out in *Analytic Britain*. We believe that if our recommendations are acted upon as a group, they will make the UK a stronger analytic nation, best placed to assume a leading role in the data economy.

GLOSSARY

- **SQL:** Relational databases to store, manipulate and query structured data
- **Basic statistics:** Basic statistical techniques for data modelling and inference, like hypothesis testing or Analysis of Variance (ANOVA).
- **Programming:** General purpose programming languages that can be applied to data manipulation and web development (e.g. C++, Java, Python).
- **Data visualisation:** Methods and tools to represent and communicate data visually, including dashboards (with tools like Tableau) and interactive data visualisations (using programming languages like D3).
- **Unstructured data analysis:** Techniques for the analysis of data that is not available in a tabular data, including text (text mining) and social networking data (social network analysis) or video (video analytics).
- **Big data:** Technologies to manage distributed and parallel data processing across multiple machines, such as Hadoop, Cassandra, Hive, Spark etc.
- **Advanced statistics:** Complex statistical techniques for data modelling and inference, including time series, non–parametric methods etc.
- **NOSQL:** Database technologies to store, manipulate and query unstructured and 'messy data' (e.g. MongoDB).
- **Machine learning:** Algorithms that automatically undertake a data task (e.g. classification or prediction), and use the outcomes of the task to improve accuracy over time (e.g. decision trees, neural networks, deep learning).
- **Experimental design:** Design of randomised control trials, A/B tests and other methods to test hypotheses via the random allocation of individuals into treatment and control groups.

APPENDIX 1 **CLUSTER ANALYSIS**

We have used a clustering algorithm to segment the companies in our sample into different Data Groups depending on their data activities. The idea here is to establish if their experience in the labour market for analytical talent varies depending on the types of data they are working with, and what they are doing with this data. We also want to explore differences in financial performance across Data groups.

1. **Select variables for clustering:** The clustering was based on company responses to three questions about data inputs and activities.
   a. Data sources used ('Data Variety'): We surveyed businesses about the regularity with which they use data from five different sources (customer data, other internal business data, social media data, 3rd party data, open data) in a scale of 1–5 (1=never, 5=routinely). For each company, we calculate the mean levels of data use across all data sources used, after removing from the sample companies that didn't provide any information about data sources (final sample for cluster analysis = 393).
   b. Data volumes used ('Data Volume'): We surveyed businesses about their data IT requirements, in a scale of 1 to 4 with 1=Excel, and 4=Multiple clusters.
   c. Use of data for decision–making ('Data Drive'): We surveyed businesses about their use of data and analysis for decision making in a scale of 1 to 5, with 1=use experience and intuition exclusively and 5=use data and analysis exclusively (this question was used to identify 'Datavores' in our Rise of the Datavores report).
2. **Imputation:** Cluster analysis requires complete cases, so we had to impute missing values to avoid losing around 1/3 of the sample in subsequent analysis. We did this using three approaches:
   a. Missing value = minimum score. This makes the (strong) assumption that companies where the respondent didn't know e.g. the data volumes being used were using the lowest volumes of data.
   b. Missing value = random score within the range of the variable. This doesn't make any assumptions about differences in the behaviours of companies with observed values and companies with missing values. It doesn't use any information from the sample in the imputation.
   c. Missing value = mean in that variable for the sector. This assumes that the missing values for companies are similar to the average in their sector. This imputation approach seems to use information from the survey effectively, so we decided to focus on it as the benchmark imputation approach going forward.
3. **Cluster selection:** Having standardised all the variables with their mean and standard deviation, we performed the cluster analysis using the K–means algorithm. To do this, one has to select the number of clusters ex–ante. To determine how many clusters to use in the analysis we looked at the changes in distance between observations and their cluster's centre as the number of clusters increases.[42] This analysis suggests there are 4–5 clusters in the data. Given the small size of the sample and the variables being used, we opted for continuing the analysis with 4 clusters. We examined their scores in the variables used for clustering and found their interpretation intuitive:

Cluster	Data Variety	Data Volume	Data Drive	Interpretation
1	+	+	++	This cluster contains businesses that are both Data-active (collect data) and data-driven (use it to make decisions)
2	–	++	–	This cluster contains businesses that collect large amounts of data
3	+	–	–	This cluster contains businesses working with many varieties (but not volumes) of data.
4	––	––	––	This cluster contains businesses that are neither data active not data driven.

4. **Cluster allocation:** K means clustering starts with a random initialisation of the clusters, which means that there may be inconsistencies in the allocation of companies to clusters across iterations. To increase the consistency of this process, we ran 1,000 iterations of the process and then allocated companies to each group using a 'majority vote' (each company went into the cluster that it had been allocated with the highest frequency).
5. **Robustness:** We compared the clustering outcomes using the 'mean' imputation approach, the minimum and random imputation approaches, as well as performing a cluster analysis with the reduced sample (only complete observations). There is a high degree of consistency between the mean, random and no imputation outcomes (companies are allocated to the same cluster 80–100 per cent of the times). There is less convergence in classification outcomes with the minimum imputation approach, although we think this is caused by the strong assumption involved in that imputation.
6. **Analysis:** We continued the analysis using the clustering outcome based on the sector mean imputation approach.

APPENDIX 2 **MODELLING**

We have a cross-section of survey data and a panel of financial performance information, mostly concentrated on the years 2009–2013. We have used this to model the link between a company's data group (based on the clustering exercise above) and its economic performance. We do this using value added (sales minus costs) as the measure of productivity, and an estimation approach with pooled Ordinary Least Square (OLS) regression with errors clustered by year/observation.

See table 3 for the results of this analysis.

- Model 2 includes the 'Data Group' predictor (using Dataphobes as a reference class), production inputs, sector and year fixed effects and interactions between production inputs and industry.

- Model 3 includes other firm-level variables including age, and measures of product and process innovation. The former comes from FAME. The last two are self reported measures of performance in a 1–5 scale which we include to try to account for unobserved heterogeneity between companies (the fact that 'innovative businesses' can be more productive and more data driven at the same time)

All the financial variables are logged, and winsorised at the 1 per cent level (i.e. replacing the top and bottom 1 per cent most extreme values in the variable with the 1st and 99th percentile).

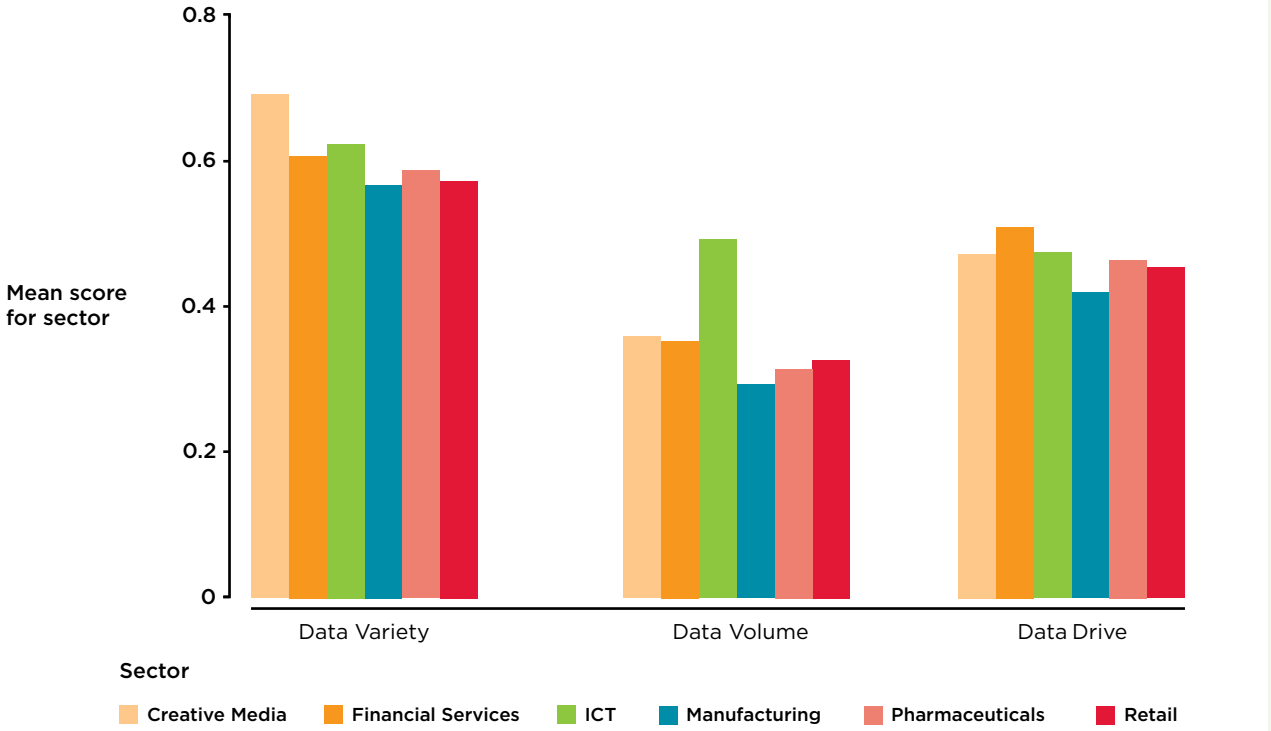
TABLE 3 **REGRESSION OUTPUTS**

	Dependent variable: Value added		
	(1)	(2)	(3)
Data Mixer	0.098** (0.047)	0.091* (0.047)	0.063 (0.048)
Data Builder	0.196*** (0.050)	0.164*** (0.050)	0.130** (0.052)
Datavore	0.210*** (0.059)	0.127** (0.060)	0.122** (0.061)
Constant	1.812*** (0.117)	1.946*** (0.220)	1.894*** (0.242)
Factor controls	Y	Y	Y
Sector controls and sector/factor interactions	N	Y	Y
Age and innovation controls	N	N	Y
Observations	1,133	1,133	1,133
R2	0.810	0.827	0.830
Adjusted R2	0.806	0.812	0.812
F Statistic	963.087*** (df = 5; 1127)	265.675*** (df = 20; 1112)	235.072*** (df = 23; 1109)

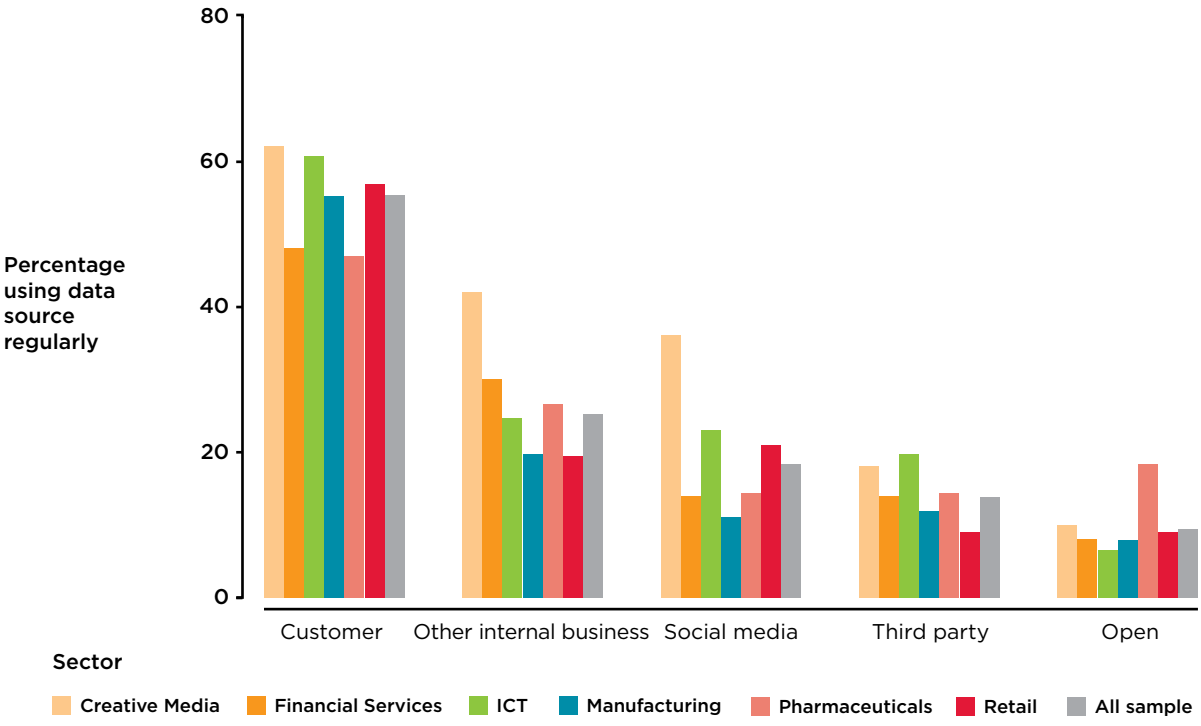
Note: *p<0.1; **p<0.05; ***p<0.01

APPENDIX 3 SECTORAL PROFILES

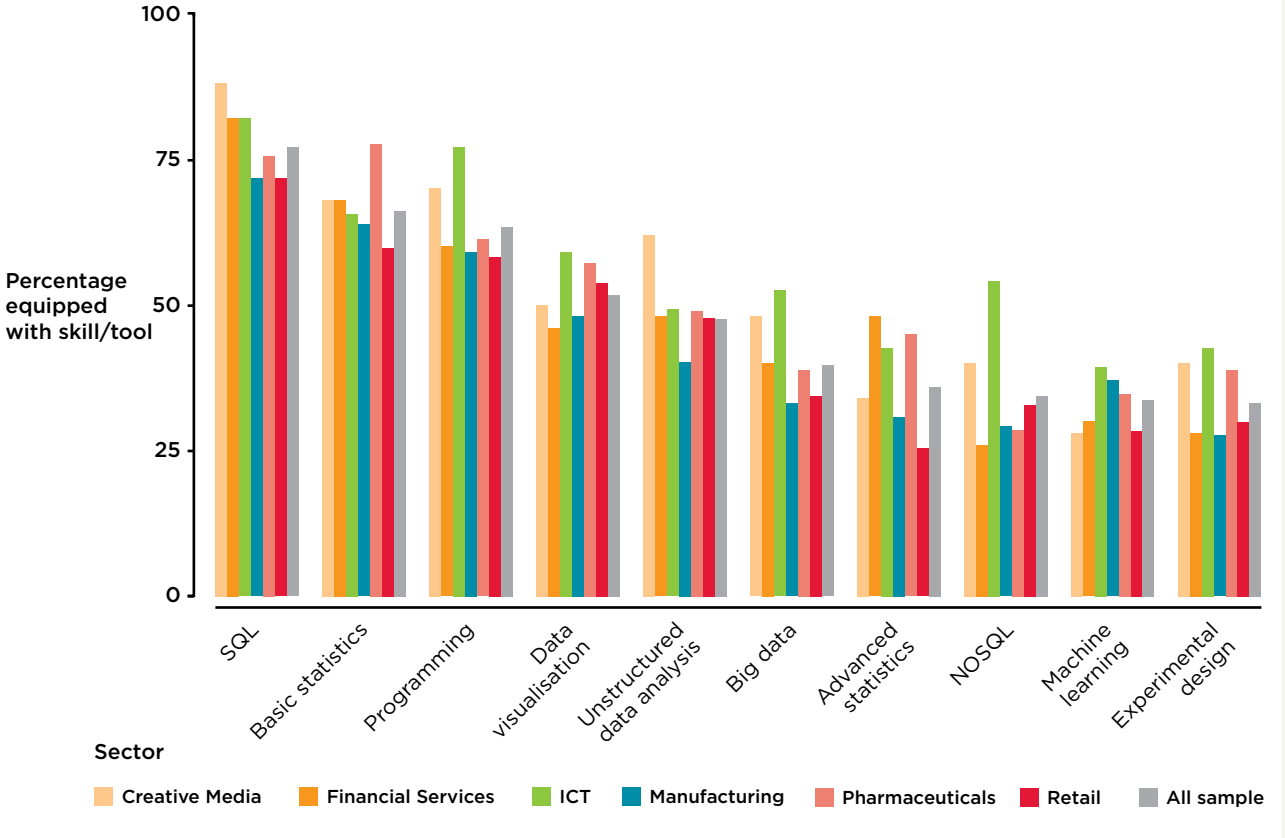
SECTOR FIGURE 1 SCORES IN DATA SOURCES, DATA VOLUMES AND DATA DRIVE MEASURES



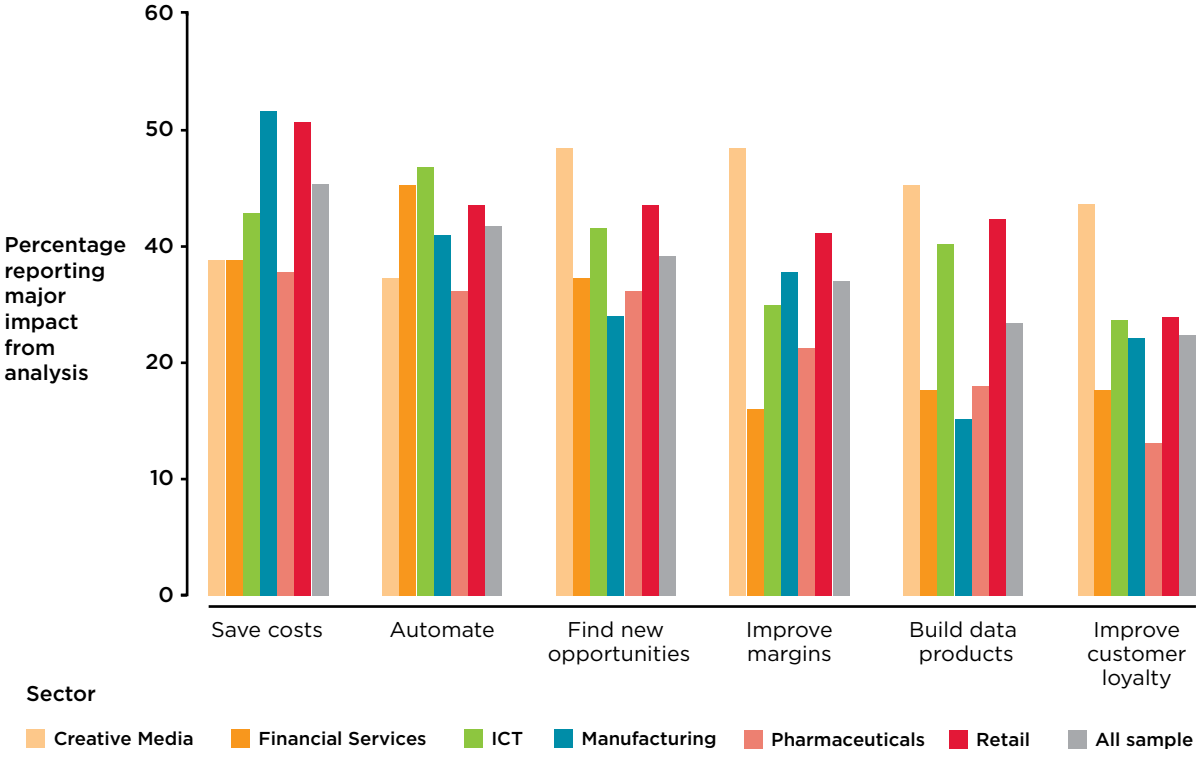
SECTOR FIGURE 2 DATA SOURCE USES



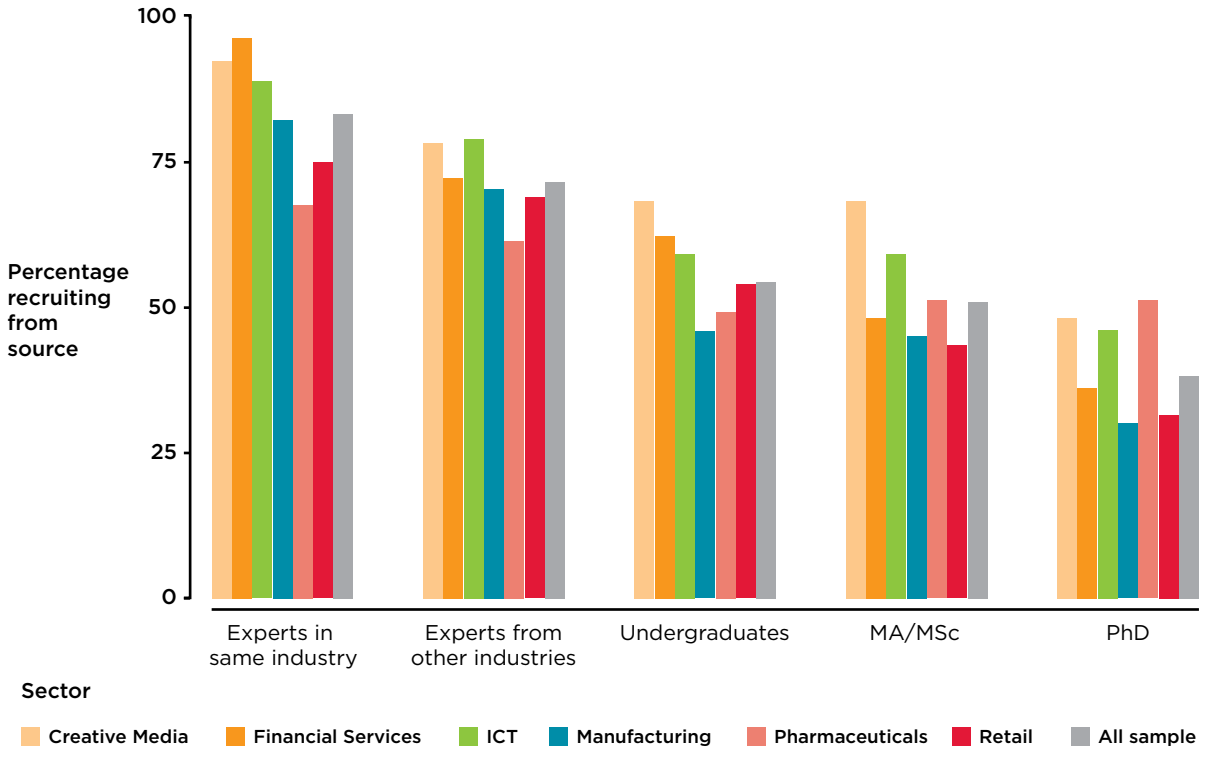
SECTOR FIGURE 3 SKILLS TO WORK WITH DATA MANAGEMENT AND ANALYTICAL TECHNOLOGIES



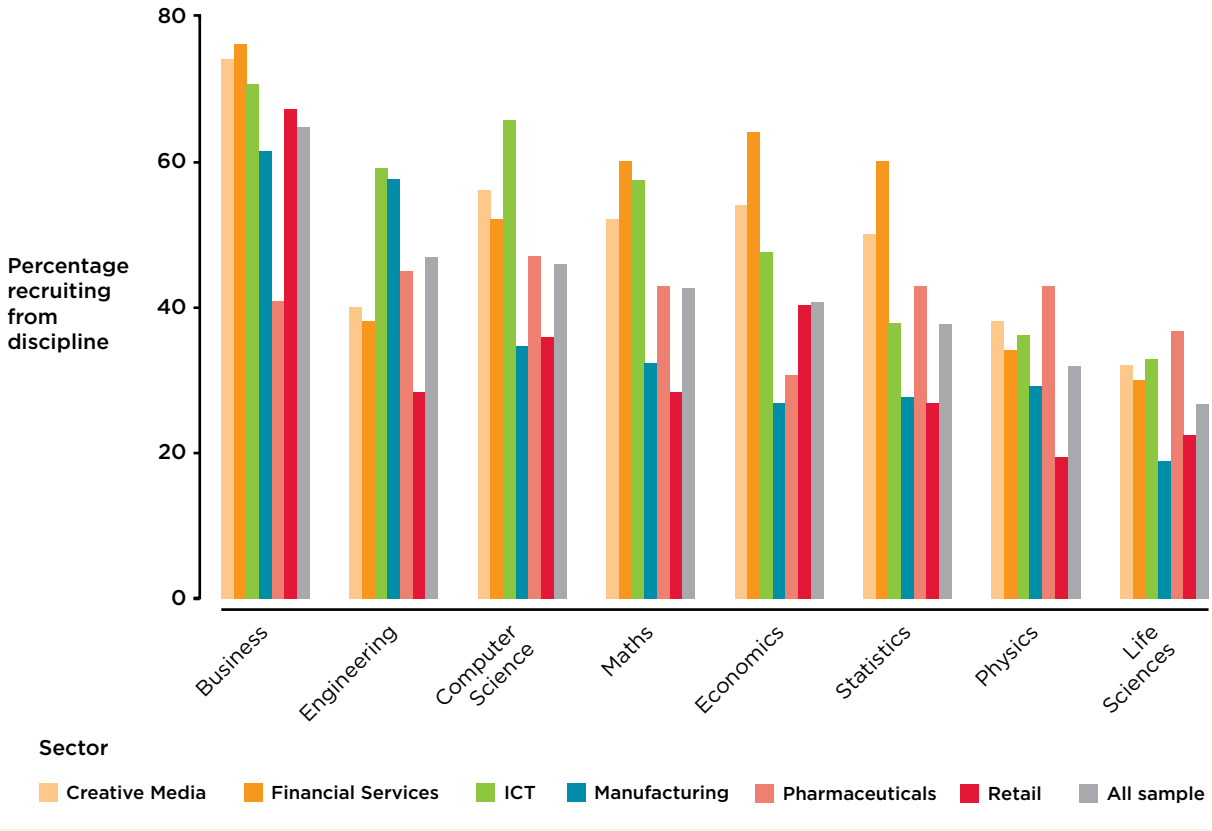
SECTOR FIGURE 4 IMPACTS OF DATA ANALYSTS



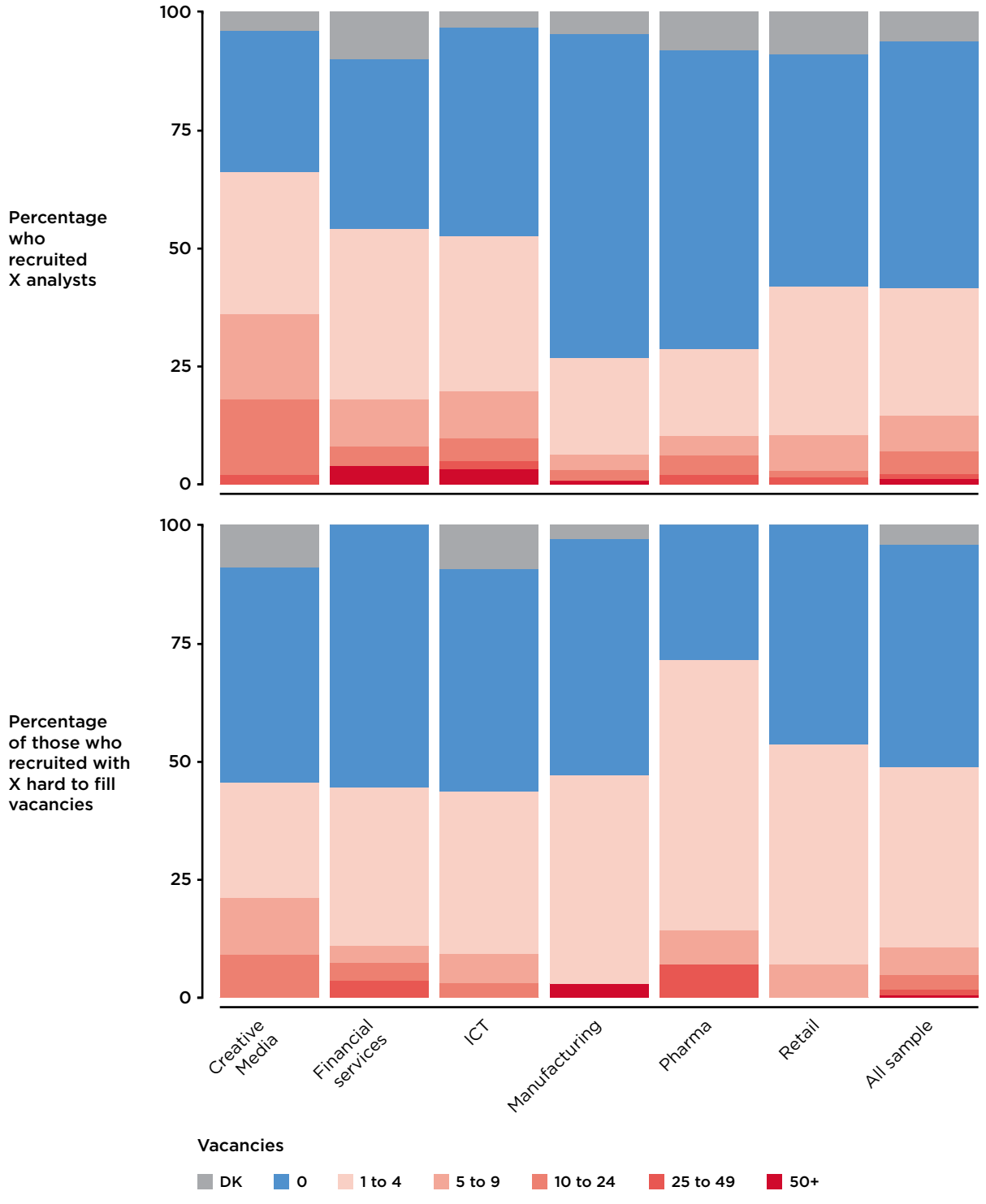
SECTOR FIGURE 5 **TALENT SOURCES**



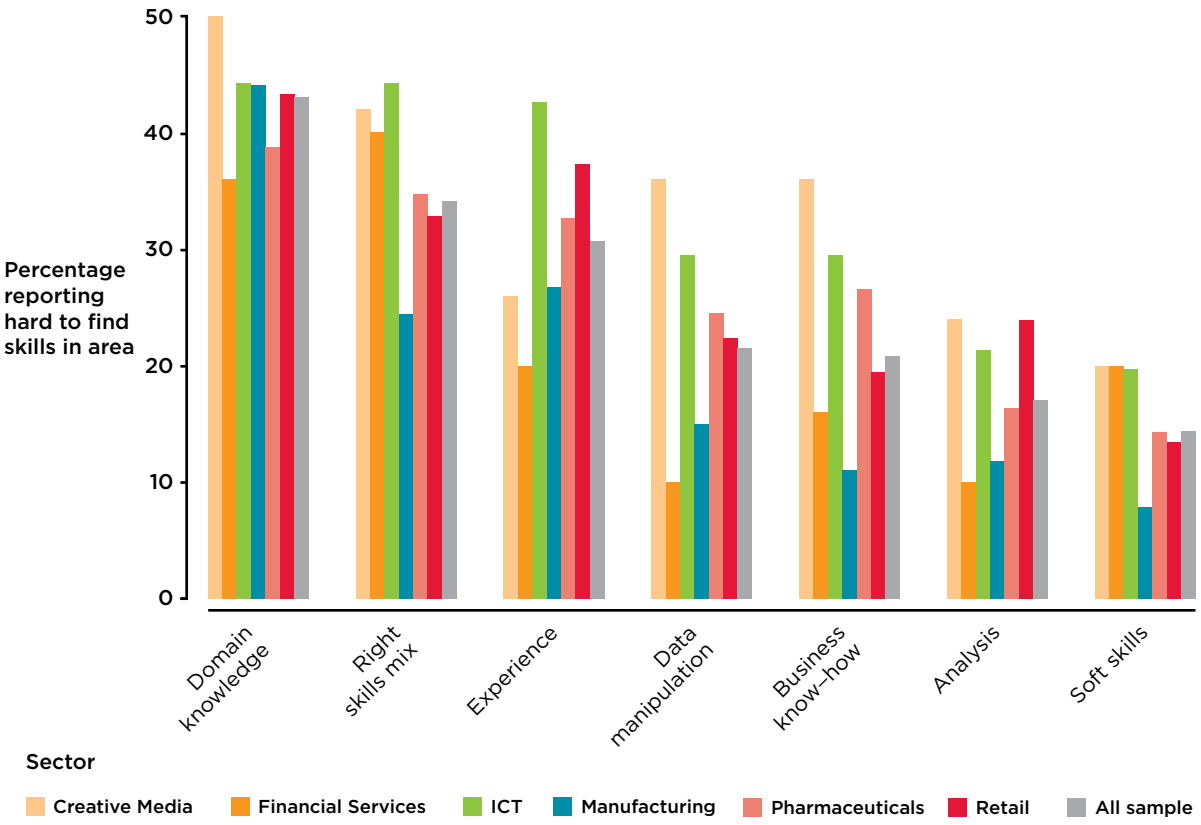
SECTOR FIGURE 6 **DISCIPLINARY SOURCES OF TALENT**



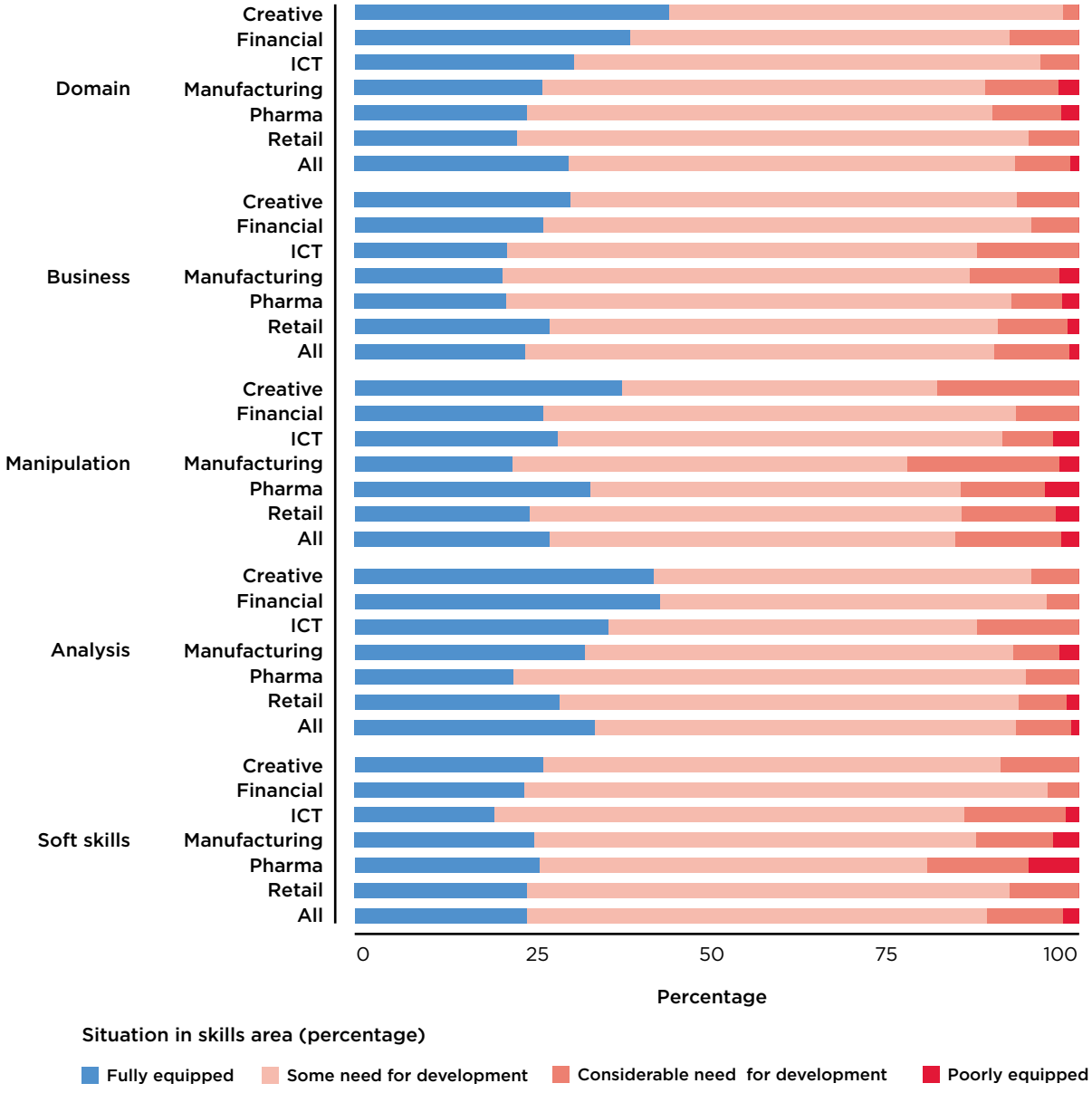
SECTOR FIGURE 7 **RECRUITMENT ACTIVITIES AND HARD-TO-FILL VACANCIES**



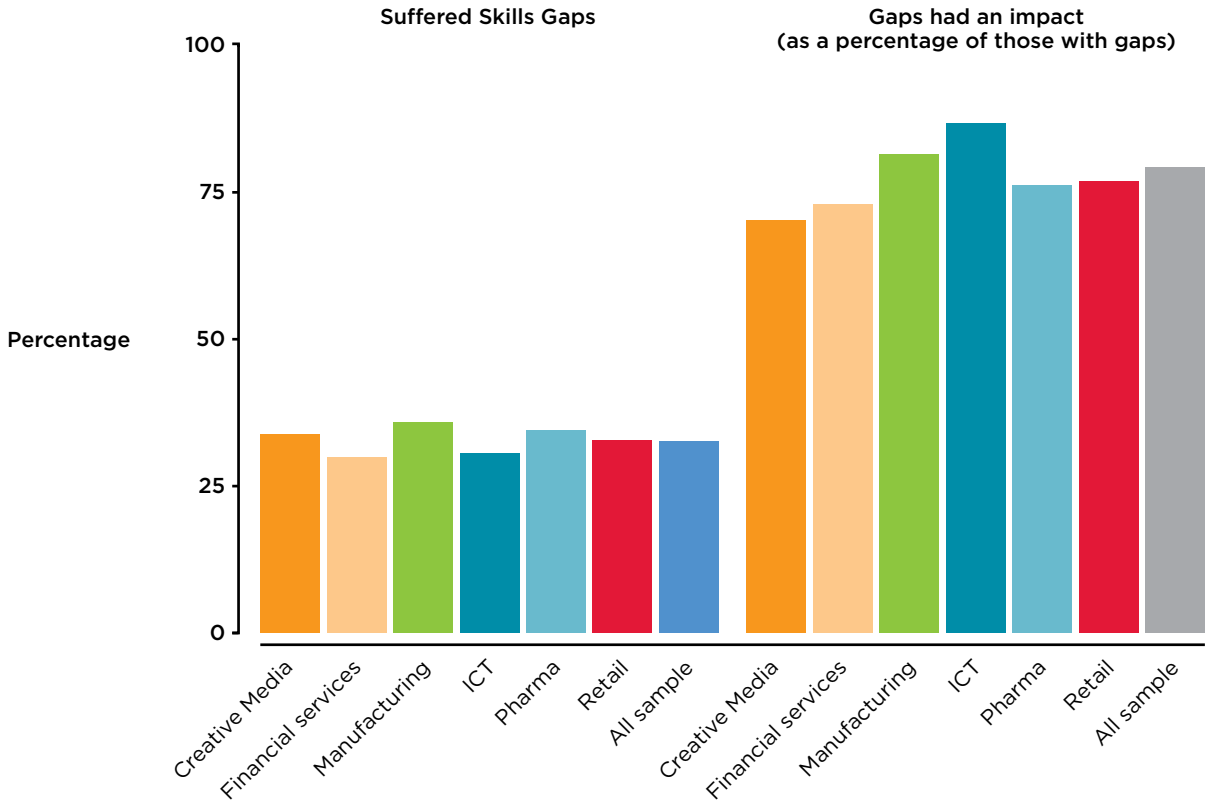
SECTOR FIGURE 8 DIFFICULTIES FINDING TALENT WITH THE RIGHT SKILLS IN DIFFERENT AREAS



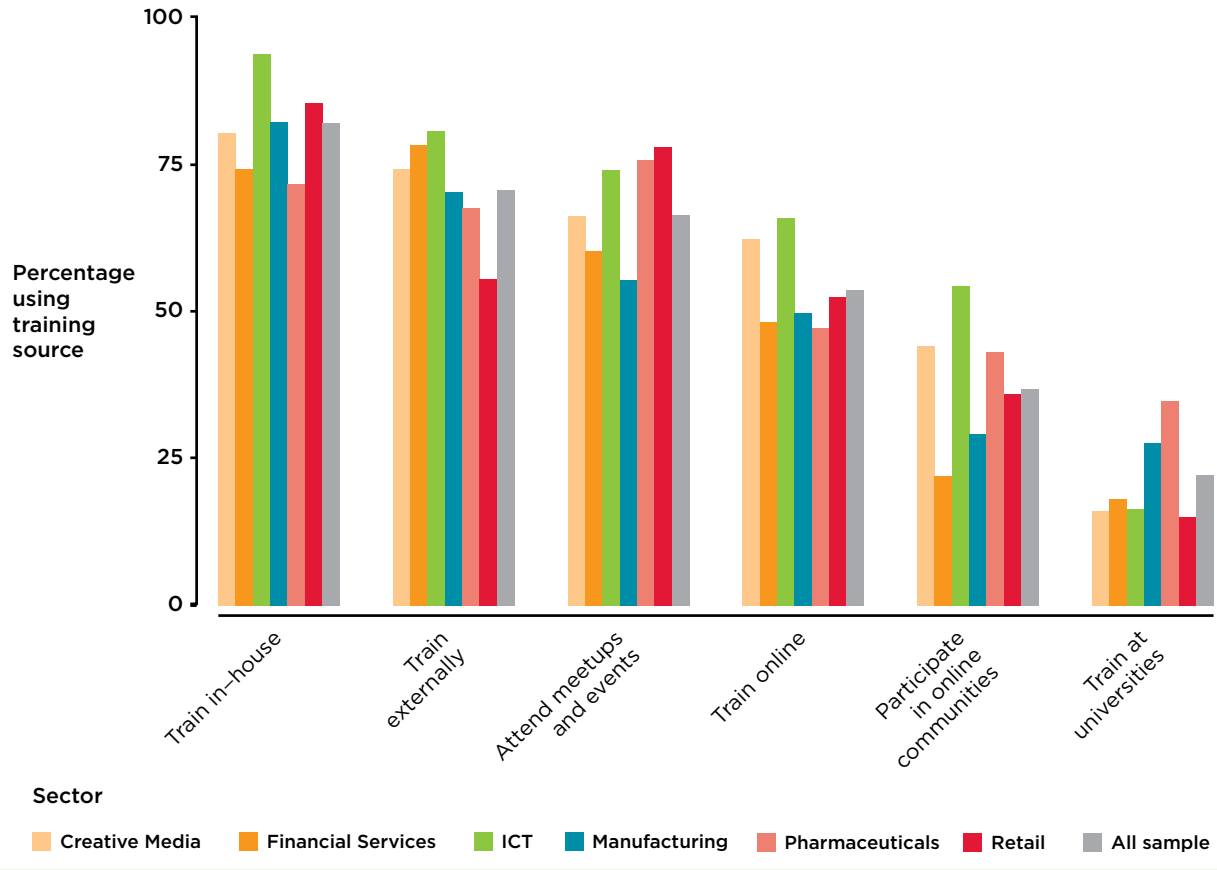
SECTOR FIGURE 9 **SKILLS DEVELOPMENT NEEDS**



SECTOR FIGURE 10 **SKILLS GAPS AND IMPACTS**



SECTOR FIGURE 11 **TRAINING ACTIVITIES AND SOURCES**



ENDNOTES

1. Bakhshi, H. and Mateos-Garcia, J. (2012) 'Rise of the Datavores: how UK businesses can benefit from their data.' London: Nesta; Bakhshi, Bravo-Biosca, A. and Mateos-Garcia, J. (2014) 'Inside the Datavores: estimating the effect of data and online analytics on firm performance.' London: Nesta.
2. http://www.cio.com/article/2854720/data-center/how-and-why-facebook-excels-at-data-center-efficiency.html
3. http://www.internetlivestats.com/google-search-statistics/ ; https://www.youtube.com/yt/press/en-GB/statistics.html
4. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
5. Mayer-Schönberger, V. and Cukier, K. (2013) 'Big Data: A Revolution That Will Transform How We Live, Work, and Think.' Boston MA: Houghton Mifflin Harcourt.
6. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
7. For more information on these technologies visit http://hadoop.apache.org/ (Hadoop), http://cassandra.apache.org/ (Cassandra) and https://hive.apache.org/ (Hive).
8. Brynjolfsson, E., Hitt, L.M. and Kim, H.H. (2011) 'Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?' SSRN Scholarly Paper. Rochester NY: Social Science Research Network. See: http://papers.ssrn.com/abstract=1819486; Bakhshi,H., Bravo-Biosca, A. and Mateos-Garcia, J. (2014) 'Inside the Datavores: Estimating The Effect Of Data And Online Analytics On Firm Performance.' London: Nesta. See also???
9. Bakhshi, H., Bravo-Biosca, A. and Mateos-Garcia, J. (2014) 'Inside the Datavores: Estimating The Effect Of Data And Online Analytics On Firm Performance.' London: Nesta.
10. SAS and Tech Partnership (2014) 'Big Data Analytics Assessment of Demand for Labour and Skills 2013-2020.' London: Tech Partnership. See: http://www.thetechpartnership.com/globalassets/pdfs/bigdata_report_nov14.pdf; Manyika, J. et al., (2011) 'Big Data: The next Frontier for Innovation, Competition, and Productivity.' McKinsey Global Institute; Booz Allen Hamilton (2013) 'The Field Guide to Data Science.' Booz Allen Hamilton.
11. Manyika et al., (2011) 'Big Data: The next Frontier for Innovation, Competition, and Productivity.' McKinsey Global Institute.
12. SAS and Tech Partnership 'Big Data Analytics Assessment of Demand for Labour and Skills 2013-2020.' London: Tech Partnership.
13. For detailed discussions about the skills and competencies of data scientists see Patil,D.J. (2011) Building Data Science Teams. 'O'Reilly Radar.' 16 September 2011. See: http://radar.oreilly.com/2011/09/building-data-science-teams.html. Conway,D. (2010) 'The Data Science Venn Diagram.' See: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram., and Schutt, R. and O'Neil, C. (2013) 'Doing Data Science: Straight Talk from the Frontline.' O'Reilly Media, Inc.
14. The aforementioned Tech Partnerhsip research suggests that Data Scientists were the hardest to recruit big data occupation.
15. Bakhshi, H., Mateos-Garcia, J. and Whitby, A. (2014) 'Model Workers.' London: Nesta.
16. Bureau Van Dijk's FAME (Financial Analysis Made Easy) is a database of companies in the UK and Ireland. It draws on business registry data (Companies House in the case of the UK) and other sources to provide financial information about nine million companies. For more information, visit: http://www.bvdinfo.com/Products/Company-Information/National/FAME.aspx
17. e-Skills UK (2013) 'Big Data Analytics Adoption and Employment Trends, 2012-2017.' London: e-Skills.
18. Manyika et al., (2014) 'Big Data: The next Frontier for Innovation, Competition, and Productivity.' McKinsey and Company. Stone, M. (2014) 'Big Data for Media.' Oxford: Reuters Institute for the Study of Journalism. See: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Big%20Data%20For%20Media.pdf; 'How Big Data Can Revolutionize Pharmaceutical R & D | McKinsey & Company,' accessed May 6, 2014, http://www.mckinsey.com/insights/health_systems_and_services/how_big_data_can_revolutionize_pharmaceutical_r_and_d.
19. Livingstone, I. and Hope, A. (2011) 'Next Gen. Transforming the UK into the World's Leading Talent Hub for the Video Games and Visual Effects Industries. London: NESTA; Bakhshi, H. and Mateos-Garcia, J. (2012) 'Rise of the Datavores.' London: Nesta; Winterbotham, M. 'UK Commission's Employer Skills Survey 2013: UK Results, January 2014.' See: http://dera.ioe.ac.uk/19271/1/evidence-report-81-ukces-employer-skills-survey-13-full-report.pdf.
20. OECD (2014) ' Data-Driven Innovation for Growth and Well-Being. Interim Synthesis Report.' Paris: OECD. See: http://www.oecd.org/sti/inno/data-driven-innovation-interim-synthesis.pdf; Bakhshi , H. and Mateos-Garcia, J. (2014) 'Rise of the Datavores.' London: Nesta.
21. This is almost identical to the proportion of Datavores in an independent firm survey of UK internet economy businesses in 'Rise of the Datavores.'
22. Each Data Group has a statistically significant higher score than the sample overall in the variable that 'defines' it. Dataphobes have significantly lower scores than the average in all of the data variables we use in our clustering.
23. http://sloanreview.mit.edu/reports/analytics-innovation/analytical-innovators/
24. Rincon, A. , Vecchi, M. and Venturini, F. (2013) 'ICTs as a General Purpose Technology: Spillovers, Absorptive Capacity and Productvity Performance.' NIESR Discussion Paper 416.
25. Nelson, R.R. (2003) On the Uneven Evolution of Human Know-How. 'Research Policy.' 32, no. 6 (2003): 909-22.
26. And with the findings of 'Rise of the Datavores.'
27. All the differences are statistically significant at the 5 per cent level.
28. They are still statistically significant for Datavores and Data Buillders (at the 10 per cent level), but they are not statistically significant for the Data Mixers.
29. Value added is defined as sales minus costs. Our analysis considers the link between being in a Data Group and a company's output conditional on its use of inputs like capital and labour. It shows the extent to which companies in different Data Groups are more or less productive for a given level of resource. Higher levels of productivity could be expected among companies with stronger economic efficiency and superior product quality allowing them to set prices higher for a given level of input.

30. Of course, our analysis cannot categorically demonstrate a causal effect of being in a Data Group on company performance. In order to do that we would need a comprehensive set of control variables or longitudinal data about levels of data activity (whereas our survey dataset is cross-sectional). This means that we cannot rule out that our findings are explained by unobserved heterogeneity (innovative, forward thinking and well-managed companies are more likely to work with data and to be more productive, for example), or reverse causality (commercially successful companies invest more in their data).
31. Several companies we interviewed in 'Model Workers' said they were looking for talent in universities in response to what they perceive to be a 'data talent crunch' for more experienced talent.
32. We have also compared perceived skills shortages between companies that sought to recruit in the previous 12 months, and companies that did not. In general, companies that tried to hire talent had a dimmer view of the situation in the labour market, highlighting skills shortages more often and in more areas than those that did not try to recruit. This is consistent with the idea that our findings reflect the reality of the analytical labour market, rather than just the perceptions of our respondents.
33. Recall this was the Data Group that faces greatest difficulties when filling vacancies.
34. By comparison, the most recent UK CES Employer Skills surveys shows that two-thirds of employers (66 per cent) had arranged or funded off – or on-the-job training or development for any of their staff in the previous 12 months. Winterbotham, M. et al., (2014) 'The UK Commission's Employer Skills Survey 2013: UK Results.' London: UK commission for Employment and Skills.
35. Levy, S. (1984) 'Hackers: Heroes of the Computer Revolution.' New York NY: Doubleday.
36. Genzel, D. (2010) 'Automatically Learning Source-Side Reordering Rules for Large Scale Machine Translation.' In 'Proceedings of the 23rd International Conference on Computational Linguistics.' Association for Computational Linguistics. 376–84 See: http://dl.acm.org/citation.cfm?id=1873824; Gai, P. Haldane, A. and Kapadia, S. (2011) Complexity, Concentration and Contagion. 'Journal of Monetary Economics.' 58, no. 5 (2011): 453–70.
37. Nelson, R.R. 'On the Uneven Evolution of Human Know-How. Columbia University.
38. Kalapesi, C., Willersdorf, S. and Zwillenberg, P. (2010) 'The Connected Kingdom: How the Internet Is Transforming the UK Economy.' The Boston Consulting Group. See: https://www.bcgperspectives.com/content/articles/media_entertainment_technology_software_the_connected_kingdom/.
39. See: Bakhshi, H. and Mateos-Garcia, J. (2012) 'Rise of the Datavores.' London: Nesta; Bakhshi, H., Bravo-Biosca, A. and Mateos-Garcia, J. (2014) 'Inside The Datavores.' London: Nesta; Bakhshi, H., Mateos-Garcia, J. and Whitby, A. (2014) 'Model Workers.' London: Nesta; Tech Partnership big data analytics reports.
40. The Tech Partnership employer skills survey, 2015.
41. See: http://www.theguardian.com/media-network/2015/feb/12/data-scientists-as-rare-as-unicorns
42. The British Academy (2015) 'Count Us In: Quantitative skills for a new generation.' Available at: http://www.britac.ac.uk/policy/count_us_in_report.cfm?frmAlias=/countusin/
43. The idea is to find the number of clusters beyond which adding more clusters doesn't substantially improve the precision of the cluster allocation.

Nesta...

Nesta

1 Plough Place
London EC4A 1DE

research@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk

www.nesta.org.uk