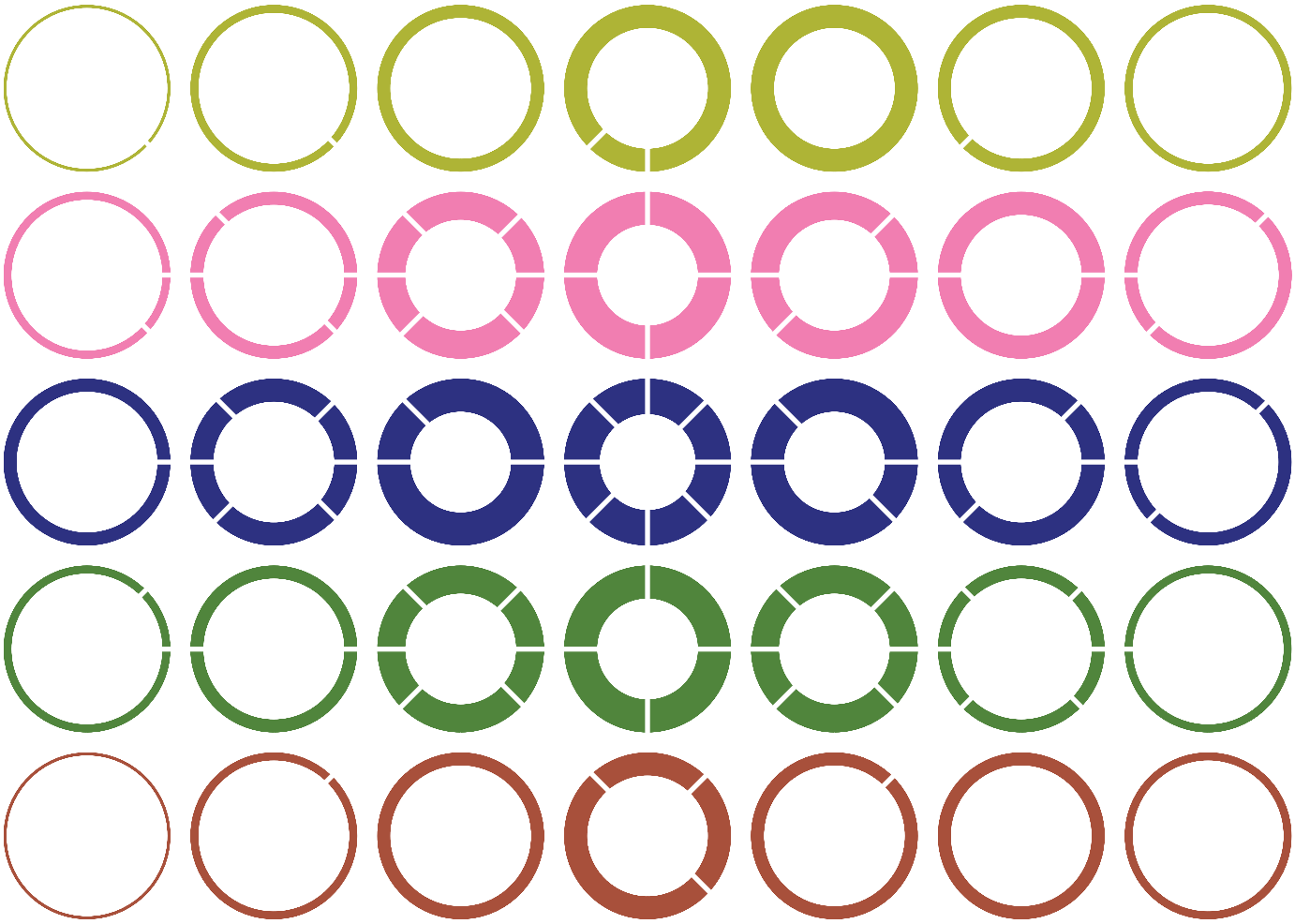
GROWTH INTELLIGENCE

National Institute of Economic and Social Research

MEASURING THE UK'S DIGITAL ECONOMY WITH BIG DATA



Max Nathan and Anna Rosso
with Tom Gatten, Prash Majmudar and Alex Mitchell

CONTENTS

FOREWORD - HAL VARIAN, CHIEF ECONOMIST, GOOGLE

Measuring better

The British mathematician and physicist Lord Kelvin famously said "If you can not measure it, you can not improve it."

In 1948, the post-war British Government badly needed to improve the nation's economy. To help it draw up the right policies for jobs and growth, the Government took a very scientific approach - one of which Lord Kelvin would have been proud. Government statisticians were instructed to count, classify and measure the economic activity of every business in the country. They developed a set of Standard Industrial Classification (SIC) codes and the data they collected was used to shape policy in every aspect of the British economy.

This was a remarkable endeavour. Here was a Government using big data to make smart decisions long before it was fashionable.

However, 65 years and several revisions later, these SIC codes are no longer adequate. Far from providing an accurate picture of the economy, they only serve to show us how much we don't know. One in ten companies in the UK are now classified vaguely as 'other'. One in five have no classification at all.

It is important not just for statisticians, economists and policymakers that we measure the economy accurately but for every one of us who are affected by the major economic decisions that Governments take based on that data.

It is clear we need a new way of measuring the economy and that is what is presented in this report by the National Institute for Economic and Social Research, based on the pioneering big data techniques of Growth Intelligence, a UK company who specialise in tracking and measuring the economic activities of companies. The richness of Growth Intelligence's data, drawn not just from official tables and accounts but gleaned from more than five billion data points, provides us with a level of detail and insight that statisticians in 1948 could scarcely have imagined.

The digital economy in the UK

Using Growth Intelligence data as a new way of measuring the economy, this report by NIESR focuses on the digital economy and provides us with a transformative view of its scale, geographical spread, revenue growth and employment.

In June 2013, the Government estimated that there were 120,000 businesses in the digital economy, based on the information available in SIC codes. Aware of the limitations of SIC codes, the Government attached a warning to this estimate: 'we may not have an exact picture of the number of businesses in the information economy, or its employment, or the value it brings to the UK economy.'

NIESR's most conservative estimate of the number of digital companies is 269,695. More than double the Government estimate. (Their most generous measure is 471,120 - almost four times the Government estimate.)

The UK is one of the world's strongest internet economies yet the myth persists that it consists largely of tiny dotcom or biotech startups in a few high technology clusters that quickly bubble up and often go bust.

The reality, as this report shows, is that the digital economy has spread into every sector, from architecture firms whose activities have become almost entirely digital to machine tool manufacturers who now use huge online data-processing facilities, such as Hadoop, to monitor every aspect of their processes.

The digital economy has spread into every part of the United Kingdom, not just in London and the South East but throughout the country, with particularly great intensity in places like Manchester, Middlesbrough and Aberdeen.

This report also shows that those firms that are part of the digital economy are, on average, growing faster than those firms that are not.

And while eminent economists on both sides of the Atlantic continue to debate whether technology and the Internet are creating or destroying jobs - driving efficiencies and productivity improvements or simply hollowing out whole types of jobs - this report shows that digital economy companies on average employ more people than those that are outside the digital economy: 23.37 compared to 20.40.

This is a groundbreaking and important report by NIESR not just because it shows that the spread of the digital economy into other sectors is driving growth and jobs throughout the UK but because - for the first time in 65 years - it presents us with a new way of measuring the economy that can only help us to take the right steps to support growth and jobs.

ABOUT THE AUTHORS

Dr Max Nathan is a Senior Research Fellow at NIESR and a Research Fellow at the Spatial Economics Research Centre.
Anna Rosso is a Research Fellow at NIESR and a Research Officer at CReAM, UCL.
Tom Gatten is CEO of Growth Intelligence.
Prash Majmudar is CTO of Growth Intelligence.
Alex Mitchell is a data scientist at Growth Intelligence.

ACKNOWLEDGEMENTS

This report is based on research funded by Google UK. Thanks to Theo Bertram, Ralf Martin, Jonathan Portes, Rebecca Riley, Chiara Rosazza-Bondibene and Stian Westlake for helpful comments and advice. Thanks also to Rosa Sanchis-Guarner for mapping, and Nick Watts for design. The views in this report are those of the authors, not the funders. Any errors and omissions are our own.

KINDLY SUPPORTED BY

Google

GROWTH INTELLIGENCE

Growth Intelligence tracks the performance and activity of all companies in the economy in real-time by analysing the digital signatures they emit. Growth Intelligence sales software automatically brings companies ready to buy together with those ready to sell. Growth Intelligence is led by Tom Gatten, an ex-BBC broadcast journalist and Prash Majmudar an ex-BAE Systems software engineer. They are turning sales from an art into a science and making tens of millions in revenue for their clients as a result. Growth Intelligence boasts Hal Varian, Google's Chief Economist as an advisor.
Visit growthintel.com to find out more.

NIESR

The National Institute of Economic and Social Research (NIESR) is Britain's longest established independent research institute, founded in 1938. The vision of our founders was to carry out research to improve understanding of the economic and social forces that affect people's lives, and the ways in which policy can bring about change. Today, we apply our expertise in both quantitative and qualitative methods, and our understanding of economic and social issues, to current debates and to influence policy. A charity, the Institute is independent of all party political interests and receives no core funding from government.
Visit www.niesr.ac.uk to find out more.

KEY FINDINGS

- The digital economy is poorly served by conventional definitions and datasets. Big data methods can provide richer, more informative and more up to date analysis.

- Using Growth Intelligence data on a benchmarking sample, we find that the digital economy is substantially larger than conventional estimates suggest. On our preferred measure, it comprises almost 270,000 active companies in the UK (14.4% of all companies as of August 2012). This compares to 167,000 companies (10.0%) when the Government's conventional SIC-based definitions are used.

- SIC-based definitions of the digital economy miss out a large number of companies in business and domestic software, architectural activities, engineering, and engineering-related scientific and technical consulting, among other sectors.

- Companies in the digital economy have a similar average age to those outside it. Shares of start-ups (companies up to three years old) are very similar. Given the popular image of the digital economy as start-up dominated, this may be surprising to some. As digital platforms and tools spread out into the wider economy, and become pervasive in a greater number of sectors, so the set of 'digital' companies widens.

- Inflows of digital companies into the economy have always been relatively small, given its sectoral share. However, using our new definitions of the digital economy, inflow levels are substantially higher.

- As far as we can tell, digital economy companies have lower average revenues than the rest of the economy, but the median digital company has higher revenues than the median company elsewhere in the economy. Revenue growth rates are also higher for digital companies. However, these results come from a sub-sample of older, likely stronger-performing companies, so there is some positive selection at work.

- Switching from SIC-based to Growth Intelligence-derived measures substantially increases the digital economy's share of employment, from around 5% to 11% of jobs. Digital economy companies also show higher average employment than companies in the rest of the economy (this reverses when we use conventional SIC-based measures of the digital economy). Looking at median employees per firm, the digital/non-digital differences are always a lot smaller. Our employment results should also be treated with some care, as not all companies report their workforce information.

- The digital economy is highly concentrated in a few locations around the UK: Growth Intelligence software provides a fresh look at these patterns. In terms of raw firm counts, London dominates the pictures, but Manchester, Birmingham, Brighton and locations in the Greater South East (such as Reading and Crawley) also feature in the top 10. Location quotients show the extent of local clustering, which for the UK's digital economy is highest for areas in the Western arc around London, such as Basingstoke, Newbury and Milton Keynes. Areas like Aberdeen and Middlesbrough also show high concentrations of digital economy activity.

01 INTRODUCTION

BACKGROUND

This report is about the digital economy in the UK. Like other parts of the 'knowledge economy', the digital economy will help shape the UK's economic prospects. Innovation is a critical factor in shaping long-term economic development; new ideas are foundational conditions for growth (Lucas 1988; Romer 1990). From a growth perspective, then, the most important parts of our industrial structure are high-skill, high-wage, highly innovative elements such as high-tech manufacturing, science and engineering, parts of financial services and parts of the creative industries. These sectors make important direct contributions to national economies; they also make critical indirect contributions via halo effects. US evidence suggests that each knowledge economy job supports up to five jobs elsewhere in the economy (Moretti 2012). And the digital economy, specifically, generates products and services that are widely used in business and in everyday life.

In principle, these are all reasons for UK policymakers to try and help the digital economy – and the wider knowledge economy – grow and thrive. Along with many others around the world, the UK government is taking a fresh look at industrial policies (Cable 2012). The Coalition's nascent industrial strategy includes the 'information economy' as one of its seven key sectors (Department for Business Innovation and Skills 2012c). The growing high-tech cluster in East London provides another focus for government policy, in the shape of the 'Tech City' strategy (Cameron 2010; Nathan 2011). The digital economy has a largely urban footprint – as we will show – and city-regions are also developing parallel economic strategies, as part of new City Deals (Cabinet Office 2012; Clark 2012).

1.1 THE PROBLEM

All of this suggests that UK policymakers – and businesses – need a good understanding of the digital economy: as a potential driver of economic growth; as potential suppliers, partners and clients; and as investment opportunities. But as it stands, the UK's digital economy is also poorly understood by researchers, policymakers and many in business. A popular view of the digital economy is that it is a) small, b) dominated by start-ups, with c) low revenue d) low employment, and e) based in a few London clusters like Tech City. This report will explore how far reality corresponds to these preconceptions.

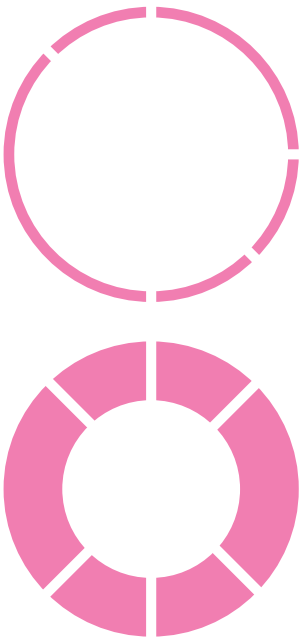
One reason why the digital economy has this popular image is that there are serious problems with both existing classifications and conventional datasets (which we discuss further in the next chapter). These mean that it has been difficult to get a proper fix on the size of the digital economy (in terms of companies, revenues and jobs generated, for example); or the nature of digital economy companies (in terms of products and services, or spatial footprint).

1.2 ABOUT THIS REPORT

This report provides new evidence, drawing on Growth Intelligence's rich data resources, on the size and nature of the digital economy in the UK. Our research suggests that much of the popular view is off the mark. We find that the digital economy a) is substantially larger than we thought, with over 40% more companies than conventional estimates; b) has about the same age profile as the rest of the economy, with the same share of start-ups among active companies; c) has lower average revenues per company than non-digital businesses, but the median digital company has higher revenues and higher revenue growth than the median company outside the digital sector; d) is likely to have slightly higher employment per company, and has a substantially higher share of national employment than conventional estimates suggest; e) is dominated by London and the Greater South East, but with clustering in Manchester, Birmingham, Bristol, Brighton, Aberdeen, Middlesbrough and other cities.

Our research is part of an on-going programme of work supported by Google UK, with further phases supported by NESTA. It contributes to a small but growing field of work using 'big data' analytics (see Einav and Levin (2013), King (2013), and Choi and Varian (2012) for overviews). As Einav and Levin point out, compared to conventional quantitative resources 'big data' is characterised by massive scale, real-time availability, lack of structure, and new angles on social and economic phenomena. This presents new challenges for researchers – but also, as we hope to show here, great potential for new insights.

02
DEFINING AND MEASURING THE DIGITAL ECONOMY



The 'digital economy' is not straightforward to define, as it is variously used to refer to a set of sectors, a set of outputs (products and services), and a set of inputs (production and distribution tools, underpinned by information and communication technologies). This means that mapping the digital economy onto industries is necessarily imprecise.

In this report we start with the UK Government's basic definition of the digital economy as comprising two sectoral groups, 'information and communications technology' and 'digital content' (Department for Business Innovation and Skills and Department for Culture Media and Sport 2009). These are cashed out using a set of detailed industry codes, or SIC codes (the full list is given in the Appendix). SIC codes are designed to represent a firm's 'principal business activity' - the activity that contributes most to value added (Office of National Statistics 2007). SICs are mainly about outputs ('the character of goods and services produced'), but also aggregate information on inputs ('inputs, processes and technologies'), and clients ('the uses to which these are put').

What do these two groups look like? Information and communications technology comprises producers of ICT systems (like broadband networks), hardware (such as computers and servers), software, and those who provide related services around these products (like sales, installation and maintenance). These sectors are characterised in terms of their outputs. Digital content is a little harder to make sense of. One recent analysis (Centre for International Economics 2005) breaks it down into 'core' content industries, where the output is digital (such as ebooks or viral videos); 'embedded' content industries, where digital content is an input into offline products and services (such as the use of computer-aided design in architecture); and distributors of digital content (such as TV and radio networks).

This still leaves some ambiguities. Some 'creative digital' firms integrate all three of these elements, working on multiple platforms and distributing their own content (Foord 2013). More seriously, it is hard to think of industries which do *not* feature digital inputs. Computers are now pervasive in most workplaces; many firms use electronic production platforms (such as the word processing software much of this report was written on), and the use of online / computerised distribution and sales platforms has also spread (such as into the retail sector).[1]

This suggests that defining digital content only in terms of inputs is not particularly useful. Instead, in this report we restrict 'digital content' to sectors where the only or principal outputs are digital products or services. For example, we exclude large parts of the architecture sector, but include firms specialising in CAD and technical drawing. By the same token, we exclude supermarkets, but include retailers whose principle offering is digital (such as digital music stores).

Note that this approach is hard to deliver using standard industry codes. Even using the most detailed SICs we are generally unable to look inside a given industry or set of sectors, to distinguish production techniques, products and services, distribution systems or customer types. This is particularly problematic for emerging sectors like the digital economy – parts of which are characterised by both digital/online and physical goods and services. Only using aggregate industry information risks creating 'false positives' – counting businesses in 'digital' sectors whose main products / services are still in the physical economy. Using the richer information available through the Growth Intelligence dataset, however, we are able to look within sectoral 'boxes'.

2.1 MEASUREMENT CHALLENGES

Researchers trying to measure the digital economy also face some practical problems. In the UK there are three principal issues.

The first is about definition. In the UK, the official 'digital economy' SIC codes do not capture the digital economy particularly well. There are some puzzling inconsistencies in what is included: 'software publishing' (SIC 58290) is included, but 'business and domestic software development' (SIC 62012) is not; 'reproduction of video recording' (SIC 18202) is included, but 'video production activities' (SIC 59112) are not. Although it clearly bears some resemblance to what most people would think of as the 'digital economy', we have not been able to find a clear written rationale for the current list.[2] Government officials are currently revisiting the code set, so some of these issues may be tidied up in the coming years.

The second issue is about data coverage. The UK's industrial mix is typically measured using large public datasets such as the Business Structure Database (BSD), which draws on administrative records. However, the BSD only includes companies paying VAT – that is, companies reporting over £79,000 in annual turnover - and/or those with at least one employee registered in the PAYE system (Office of National Statistics 2010). As such, the BSD will miss out a large number of start-ups and early stage companies. Given the preponderance of such start-ups and young companies in emerging sectors, it is likely that many digital companies will be missed.[3]

The third issue is about the quality of SIC coding itself. SIC coding is often incomplete or uninformative in the data, or even missing. The incompleteness issue is particularly pronounced in datasets that get around the coverage issue, such as Companies House (which covers all registered companies in the UK, and is described in more detail in section A1, in the Appendix). Ideally,

[1] Notably, the CIE analysis includes all of the public sector as an 'embedded' digital content industry.

[2] The closest we have found is a brief treatment in the Impact Assessment report for the 2010 Digital Economy Bill. Even this gives no clear rules (BIS and DCMS 2010). Department for Business Innovation and Skills, Department for Culture Media and Sport, et al. (2010). Impact Assessment for the Digital Economy Bill. London, BIS.

[3] The Business Population Estimates dataset (BIS, 2012b) provides high-level coverage of all UK firms, but with some key caveats which make it unsuitable for use here. Only high-level UK and regional aggregates are presented, industry codes only run to SIC3 (not detailed enough for digital economy work) and unified microdata is not available to researchers (requiring separate applications through the Secure Data Service and the HMRC Data Lab).

statisticians would allocate a SIC code based on detailed information about organisations' business models and their contribution to value-added. However, individual businesses actually pick their own SIC code and are responsible for keeping it up to date.

When companies register at Companies House, for example, they are required to assign themselves a SIC code. As closely related activities fall into different SIC codes, the process is not user-friendly, and this has a number of unfortunate consequences. For instance 9.2% of companies who file at Companies House describe themselves in uninformative SICs ('Other business support service activities not elsewhere classified ' or 'Other service activities not elsewhere classified '). Further, there is no clear incentive for a company to classify themselves correctly or to amend it as they change. This means that many thousands of companies are likely to have misleading SIC codes. Worst, a large number of companies (over 20% of our raw sample) end up providing no SIC information at all.

Taken together, these three problems present profound challenges for policymakers and business analysts trying to understand the size and characteristics of the digital economy. In turn, this makes it harder to account for its economic impacts, define effective policies, or assess potential markets.

Our research uses richer data sources and definitions to work around these problems. First, we use data from Growth Intelligence, which gives us a 'benchmarking sample' drawn from all active companies in the UK. Second, we use Growth Intelligence's machine-learnt sector and product classifications to provide a cleaner definition of the digital economy.

CASE STUDY 1 - Communique

Communique Publishing publish brochures, provide marketing consultancy, communication and web design, e-Marketing and QR coding services. Its digital services include web page design and hosting, domain registration, search engine optimization and on-screen presentations.

Their SIC code is 58190 – 'Other publishing activities'. This means this fast-growing company from East Yorkshire are outside the digital economy defined by traditional methods. Growth Intelligence classified the Communique Publishing as a publisher, but the company's product as Custom Software Development, recognising their role in the digital economy.

Over the last 18 to 24 months, the web hosting and search optimization businesses took off. Simon, CEO of Communique says it was very difficult to find a SIC code represents his business. The fast evolution of Communique's services did not correspond with update to the SIC Code system. The company believes that it is very important to have a correct SIC selection, especially when trading with continental Europe, because it is a direct and easy way to describe what they do.

"I could not find a SIC code that correctly represent the business"
Simon, CEO, Communique Publishing

2.2 THE GROWTH INTELLIGENCE DATASET

Our core analytical tool and source of data is the Growth Intelligence dataset, and is a cross-section of all active companies in the UK up to August 2012. Growth Intelligence covers companies registered at Companies House, which gives information on companies' start dates, directors and shareholders, full registered addresses and balance sheets (see Appendix, section A1). Growth Intelligence software enriches this information by gathering structured information from the Internet, as well as matching in data from other public resources (such as patents information). Growth Intelligence software automatically analyses the language used by a company and by others about a company to provide sectoral and product typologies. These 'events' can be either observed (i.e. Companies House filings, patents) or derived from free text (via press or online conversations). Words, phrases and terms associated with each company (called tokens) are linked to each company. Using machine-learning algorithms developed by the company, companies are then classified into sectors and product categories. The Growth Intelligence classifications include 145 different sectors and 38 principle product groups, resulting in a potential classification of companies into 5510 (=145*38) categories (a finer classification than SICs, which group companies into a possible 806 categories). More information on the dataset is provided in section A2 of the Appendix.

2.3 CREATING A 'BENCHMARKING SAMPLE'

Our raw data comprises 3.07m UK-registered companies. We strip some of these companies out of the data in order to create a sample that allows us to compare SIC-based definitions of the digital economy against alternatives developed using Growth Intelligence categories. We call this our 'benchmarking sample'. We take a conservative approach in defining the benchmarking sample. First, we only consider those companies where a SIC code is provided[4] and drop all companies that are dormant, non-trading or insolvent[5]. Second, we drop holding companies (defined by SIC). Third, we control for the existence of clusters of related companies within the same firm. To use company name, industry code and address information, we identify companies belonging to the same group, and keep only the unit reporting the highest revenues (as we discuss later, this is likely to undercount companies' true revenue levels). After cleaning, our benchmarking sample includes 1.868m UK companies, active as of August 2012. More information on the process is provided in section A3.1.

Our approach is still likely to undercount the real size of the digital (and non-digital) economy, for two reasons. First, we only include companies who provide SIC codes, in order to benchmark conventional definitions of the digital economy against definitions based on Growth Intelligence's classifications. Second, by using registered companies, we undercount the number of business units 'on the ground' – since many companies have more than one office or branch.

[4] Dropping companies with no SIC information allows us to benchmark our estimates of the digital economy against conventional SIC-based estimates.

[5] A dormant company is a company that had no "significant accounting transactions" during the accounting period (http://www.companieshouse.gov.uk/about/gbhtml/gp2.shtml#ch8). Those who are dissolved, insolvent or have had notices of insolvency are dropped from the sample.

2.4 DEFINING THE DIGITAL ECONOMY USING GROWTH INTELLIGENCE DATA

The UK government's official definition of the digital economy uses SIC codes, which as we have discussed, may lead to errors or uninformative results. We use Growth Intelligence's multi-dimensional sector and product classifications to provide a more accurate definition based on companies' overall industry environment (sector context) and their most likely main products/services[6]. This allows us to look inside sector 'boxes' to identify companies in broadly 'digital' sectors whose main outputs are also digital. In this way we get a much cleaner sense of the digital economy, which follows the principles set out at the start of this section.

We build our new definition in six stages. In essence, we start with a 'rough cut' of the digital economy using SIC codes, look at the corresponding Growth Intelligence sector and product groups, then map sector-by-product groups back onto our benchmarking sample.

First, we take the sub-sample of companies who have digital economy SIC codes, as set out in the official BIS-DCMS list. (see Table A1 in the Appendix) We then take the corresponding Growth Intelligence relevant sectors and product classifications for those companies: this provides an initial shortlist of 21 'digital' sectors and 15 product groups. Note that these correspond exactly to the existing 'official' BIS-DCMS list. Second, we use a threshold rule to exclude sparse sectors that have marginal presence in the digital economy. Third, we manually edit the Growth Intelligence sector list to remove irrelevant sector groups. We follow our definition rules to do this, and exclude sector groups such as 'oil and energy', ' non-profit organization management', 'apparel and fashion'. We also add in some sectors where the key outputs are likely digital, such as 'on-line publishing'. Fourth, we repeat the manual edit for our product/service list, excluding physical outputs (e.g. 'consumables', 'care or maintenance', ' clothing') and including some additional digital ones (e.g. 'software desktop or server'). Fifth, we make a precision check to look inside Growth Intelligence sector and product groups. Finally, we use our clean sector-by-product categories on the full benchmarking sample.

More information on the process is provided in section A3.2 in the Appendix: the resulting list of digital sectors and digital products are set out in Tables A2 and A3. Note that this procedure has a number of attractive features. It identifies companies that are in official / conventional definitions of the digital economy, but also exploits the power of the Growth Intelligence dataset to capture those companies that would not be visible using the official industry codes. It also defines companies in terms of their industry and their principal activity, so that we exclude companies who might be in (say) publishing but operating only using physical print.

Finally, while some manual cleaning is inescapable, this is rule-driven and has the overall effect of taking away sector/product groups, rather than artificially inflating the digital economy. This implies that (for instance) if we find that the digital economy is larger than conventional estimates suggest, this will be driven by underlying features of companies (as captured in better data) rather than by simply classifying a wider set of businesses as 'digital'.

[6] At this stage of our analysis we use overall sector context and product type. In future versions of the research we will also build in Growth Intelligence's sales process and client type dimensions.

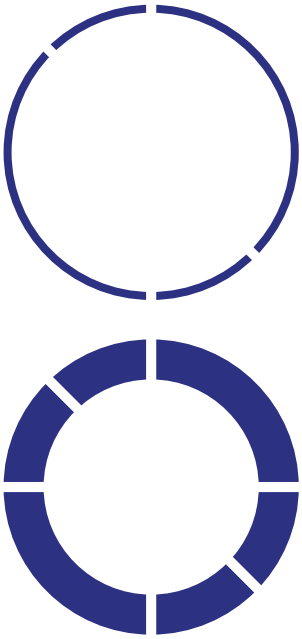
03
HOW BIG IS THE DIGITAL ECONOMY?

This section provides basic information on the size of the digital economy, comparing estimates derived from Growth Intelligence measures with estimates using conventional SIC-based measures.

3.1 FIRM COUNTS

Our benchmarking sample contains just under 1.87m active companies registered in the UK as of August 2012. Table 1, below, reports the number of companies inside and outside the digital economy, measured using conventional SIC-based definitions and with our Growth Intelligence-derived classifications. The table also sets out these sectors' shares of all companies in the sample.

In Panel A, we define the digital economy using the official classification based on SIC codes (see Table A1). We find that it represents 10.04% of the economy (187,616 companies). The digital economy can also be broken down into the information and communication technology (ICT) sectors (34.8%) and digital content industries (65.1%).



▼ **TABLE 1:** DIGITAL ECONOMY COUNTS

	OBSERVATIONS	%
A. SIC 07		
OTHER	1,681,151	89.96
DIGITAL ECONOMY	187,616	10.04
OF WHICH:		
ICT SECTOR	65,383	34.85
DIGITAL CONTENT	122,233	65.15
B.GI SECTOR AND PRODUCT		
OTHER	1,599,072	85.57
DIGITAL ECONOMY	269,695	14.43
TOTAL	1,868,767	100

Note: Other includes all the other firms. ICT sector and Digital Content follow the BIS (2009) definition.Panel B defines the digital economy using GI digital sector by digital product "cells".

▼ **TABLE 2:** DIGITAL ECONOMY COMPANIES BY SECTOR

			REVENUES	
	OBSERVATIONS	%	MEAN	MEDIAN
ANIMATION	15	0.01	2,523,751	223,217
ARCHITECTURE	47,582	17.64	1,030,438	119,438
AVIATION_AEROSPACE	371	0.14	10,788,721	990,689
BROADCAST_MEDIA	1038	0.38	4,998,423	64,515
BUSINESS_SUPPLIES_EQUIPMENT	226	0.08	8,639,873	1,087,050
COMPUTER_GAMES	33	0.01	105,300,000	3,246,058
COMPUTER_HARDWARE	264	0.1	11,189,817	403,386
COMPUTER_NETWORKING	374	0.14	1,892,677	345,218
COMPUTER_NETWORK_SECURITY	99	0.04	5,833,799	1,537,926
COMPUTER_SOFTWARE	2,046	0.76	6,672,728	1,147,716
CONSUMER_ELECTRONICS	181	0.07	80,516,082	656,124
CONSUMER_SERVICES	68	0.03	5,046,384	708,848
DEFENSE_SPACE	57	0.02	43,039,921	6,746,209
DESIGN	759	0.28	2,396,744	528,645
E_LEARNING	78	0.03	1,484,175	150,385
ELECTRICAL_ELECTRONIC_MANUFACTURING	17,704	6.56	3,622,348	123,549
ENTERTAINMENT_FILM_PRODUCTION	12,413	4.6	2,342,001	10,011
FINANCIAL_SERVICES	2,968	1.1	13,981,126	425,645
GRAPHIC_DESIGN	1914	0.71	2,572,763	429,291
INFORMATION_SERVICES	167	0.06	3,269,576	605,142
INFORMATION_TECHNOLOGY	106,193	39.38	771,513	51,854
INTERNET	1,189	0.44	2,642,974	417,799
MARKETING_ADVERTISING	11896	4.41	1,824,341	35,151
MEDIA_PRODUCTION	413	0.15	3,602,377	679,424
MEDICAL_EQUIPMENT	143	0.05	12,270,692	2,488,854
MOTION_PICTURES_FILM	338	0.13	2,171,519	750,448
MUSIC	412	0.15	2,885,209	716,595
NANOTECHNOLOGY	9	0	2,570,054	1,496,065
NEWSPAPERS_JOURNALISM	52	0.02	24,263,269	1,890,005
ONLINE_PUBLISHING	577	0.21	1,967,611	549,516
PHOTOGRAPHY	4362	1.62	367,539	29,034
PRINTING	12,576	4.66	2,101,859	184,607
PUBLISHING	3,117	1.16	4,786,819	69,285
SEMICONDUCTORS	8615	3.19	1,537,069	0
TELECOMMUNICATIONS	29100	10.79	3,320,249	24,278
WIRELESS	2346	0.87	418,384	13,263
TOTAL	**269,695**	**100**	**1,888,099**	**77,105**

Note: Observations by sector when defining digital economy using Growth Intelligence product and sector.

▼ **TABLE 3:** DIGITAL ECONOMY COMPANIES BY PRODUCTS/SERVICES

			REVENUES	
	OBSERVATIONS	%	MEAN	MEDIAN
BROADBAND_SERVICES	12,383	4.59	2,626,414	10,011
CONSULTANCY	175,238	64.98	1,316,291	73,210
DIGITAL_MEDIA	94	0.03	4,847,376	401,590
ENTERTAINMENT_EXPERIENCE_OR_ACTIVITY	836	0.31	8,072,666	672,008
ELECTRONICS	16,655	6.18	6,437,461	158,126
HARDWARE_TOOLS_MACHINERY	16,309	6.05	2,761,352	412,253
MEDIA_DISTRIBUTION	1,902	0.71	731,505	9,084
MEDIA_DELIVERY	12	0	699,077	5,361
MEDIA_PRODUCTION	2,884	1.07	3,594,331	35,858
PHOTOGRAPHY	4,652	1.72	362,658	35,437
PRINTING_SERVICES	9,955	3.69	1,754,853	145,661
PEER_TO_PEER_COMMUNICATIONS	7,244	2.69	5,724,270	45,379
SOFTWARE_DESKTOP_OR_SERVER	41	0.02	19,144,136	586,768
SOFTWARE_MOBILE_APPLICATION	32	0.01	2,299,687	1,300,314
SOFTWARE_WEB_APPLICATION	20,472	7.59	435,966	32,384
CUSTOM_SOFTWARE_DEVELOPMENT	752	0.28	4,796,575	488,865
WEB_HOSTING	234	0.09	3,371,184	215,510

Note: observations by product when defining digital economy using Growth Intelligence digital product and digital sector.

In Panel B of the table we use our preferred and more comprehensive definition of digital economy, using Growth Intelligence classifications, where we include both sectors and products as explained in the previous section. Measured this way, the digital economy now comprises 269,695 companies, around 14.4% of the economy. That is, our estimates of the size of the digital economy are over 40% higher when we move beyond SIC-based definitions.

It is worth setting out what might be driving these results. Our identification procedure aims to deal with holes in the official definition of the digital economy, and errors in SIC coding. By letting the results fall out of the information in the Growth Intelligence dataset, we are able to show the digital economy is bigger than estimated by the SIC definition. We might worry that manual reclassification is driving the shift. However, when we repeat the procedure without any manual cleaning, we find 471,120 digital economy companies, or 25.2% of the sample – a difference of almost 10 percentage points. So the net effect of our manual steps is to reduce the size of our new digital economy estimates and not to increase it.

Finally, remember that this is a benchmarking sample, where we are excluding all companies from the sample for which SIC information is not available. That is, none of the difference between conventional SIC-based estimates and our new estimates is driven by the coverage problem [7].

[7] Unless digital economy businesses are significantly less likely to report any SIC information than non-digital companies. We explore this issue further in future versions of the research.

▼ **TABLE 4:** IN WHICH SIC CODE ARE THE ADDITIONAL FIRMS WE ARE CAPTURING USING GROWTH INTELLIGENCE SECTOR AND PRODUCT DEFINITIONS?

		A		B	
DESCRIPTION	SIC2007	OBSERVATIONS	%	OBSERVATIONS	%
NON-SPECIALISED WHOLESALE TRADE	46900	1,112	1.0	1,170	0.6
VIDEO PRODUCTION ACTIVITIES	59112	51	0.0	3,379	1.6
READY-MADE INTERACTIVE LEISURE AND ENTERTAINMENT SOFTWARE DEVELOPMENT	62011	2,400	2.0	2,403	1.2
BUSINESS AND DOMESTIC SOFTWARE DEVELOPMENT	62012	17,948	15.3	17,999	8.7
COMPUTER FACILITIES MANAGEMENT ACTIVITIES	62030	2,151	1.8	2,157	1.0
OTHER INFORMATION SERVICE ACTIVITIES N.E.C.	63990	7,318	6.2	7,324	3.5
ARCHITECTURAL ACTIVITIES	71111	8,566	7.3	8,670	4.2
ENGINEERING DESIGN ACTIVITIES FOR INDUSTRIAL PROCESS AND PRODUCTION	71121	2,951	2.5	2,952	1.4
ENGINEERING RELATED SCIENTIFIC AND TECHNICAL CONSULTING ACTIVITIES	71122	5,052	4.3	5,056	2.4
OTHER ENGINEERING ACTIVITIES	71129	13,193	11.2	13,243	6.4
TECHNICAL TESTING AND ANALYSIS	71200	3,154	2.7	3,159	1.5
MEDIA REPRESENTATION SERVICES	73120	2,755	2.4	2,771	1.3
OTHER BUSINESS SUPPORT SERVICE ACTIVITIES N.E.C.	82990	4,225	3.6	4,503	2.2
OTHER SERVICE ACTIVITIES N.E.C.	96090	205	0.2	57,225	27.6
REPAIR OF MACHINERY	33120	2,044	1.7	2,048	1.0
REPAIR OF ELECTRICAL EQUIPMENT	33140	2,030	1.7	2,046	1.0
TOTAL		**74,043**	**64.01**	**134,935**	**65.64**

Note: Firms in the digital economy (GI definition) but not in the SIC code definition. The percentage refers to the percentage of firms in each SIC code excluded from the official definition (only the most relevant are reported). In Panel A, the digital economy is defined using Growth Intelligence sectors and products. In Panel B, the digital economy is only defined using Growth Intelligence sectors

More importantly, our estimates also suggest that conventional definitions and datasets tend to substantially undercount the newer parts of the economy; and conversely, that digital platforms, products and services are spreading far beyond the digital 'core' of popular imagination, across a much broader set of companies. We return to this point in Section 4.

We also provide some breakdowns of company types within the digital economy, by sector and by product group. Table 2 shows a wide spread among the 36 sector types. Around 40% of companies are located in information technology or computer-related sector groups, with a further 10.8% of companies in telecommunications. The rest are spread out in different groups like architecture (17.6%), electric and electronic manufacturing (6.5%) or printing (4.6%). Note that while many of these sector categories might otherwise include businesses producing 'non-digital' products and services, the figures here represent those who also pass our product/service criterion. Table 3 shifts the focus to principal products/services, and sets out the corresponding list. Most of these companies are providing some kind of a consultancy service (just under 65%), offering ICT-related hardware/tools/machinery (around 6.05%), software web applications (just over 7.5%) or electronics (just under 6.2%).

3.2 COMPARING NEW AND OLD MEASURES: WHO'S LEFT OUT?

Our sampling method gives us a large number of companies that we would not treat as digital economy when using only SIC codes. We map these companies back onto their relevant SIC codes in order to check what the conventional industry classification misses.

In Table 4 we report the most relevant SIC codes (defined by the percentage of companies in each industry) that are considered digital according to our measure. Around 60% of the companies end up in one of the 16 categories. In Panel A of the table we show the categories when using our preferred definition of digital economy, which interacts sectors and products. In Panel B, we show the same SIC categories but when only digital sectors in Growth Intelligence classification are used to classify digital companies. Starting from Panel A, it is interesting to notice that around 17% of the omitted companies develop software (SIC codes 62011 and 62012), which is very likely to be an important activity in the digital economy. The other companies considered are mainly doing either engineering (around 18% including codes 71121-71129) or architectural activities (around 7.3%). While we observe a large proportion of the companies classified under the sector 'other service activities' in Panel B, this sector has negligible presence in Panel A (when the digital product/service criterion is introduced). Conversely, using only sector information undercounts the presence of software firms (around 10% of observations in panel B, versus 15% in Panel A). These differences highlight the point made in section 2: it is important to use both overall sector context and detailed product/service information.

CASE STUDY 2 - Kelton Engineering

Incorporated in 1991, Aberdeen-based Kelton Engineering Limited provides 'oil flow measurement' software to the Oil and Gas industry. Clients include Shell, Mobil, and BP. Kelton is currently classified under the SIC Code 82990 – 'Other business support service activities n.e.c.'. Growth Intelligence classified the company in the Oil and Energy sector but recognizes this main product as 'Custom Software Development'. This is an example of a company deeply involved in software development and training, but completely invisible when sorting by SIC Code.

CASE STUDY 3 - Rok Operations

Incorporated in 2004 in Staffordshire, Rok Operations provides server-based text-to-speech technology to website owners. The technology allows visually impaired website users to hear the website text in their native language or choose a real-time spoken translation. Growth Intelligence classifies Rok Talk in the Translation sector, offering Software Web Application and tagged the company to 'text-to-speech', 'Web accessibility', and 'Human-Computer interaction' tokens. The company's SIC Code is 82990 - 'Other business support service activities not elsewhere classified', meaning you won't find this company looking for digital businesses by SIC code.

04
WHAT KIND OF COMPANIES?

This section provides more detailed information on digital economy companies' age, revenues and employment. Not all companies report revenue or employment data, so this analysis is done on suitable sub-samples. While some companies have no revenue or employees to report, there are also some holes in the Companies House data.[8] We perform a range of diagnostic checks to make sure the sub-samples are representative, and make use of 'nowcasted' revenue information provided by Growth Intelligence that imputes turnover in real-time using the real-time data 'signature' of a business. Even so, data limitations mean that revenue and employment information has to be interpreted with some care.

4.1 AGE

We look at the age of companies in the digital and in other sectors, defining it as the years of activity since the company was incorporated. In Table 5 we report the average age by sector for each of the different definition used. Comparing average age using SIC codes, the digital economy companies are on average 8.6 years old, almost two years younger than companies in other sectors. As previously explained the digital economy is then broken down into ICT companies and digital content companies, which are approximately the same average age. Interestingly, when using the Growth Intelligence digital economy sample (sector by product), we find that the average age of companies is 9.8 years, against 10.2 in the rest of the economy. In this case, the average digital economy company is only slightly younger.

[8] Some companies will not file annual returns or accounts on time; others may file incomplete information; others may fail to declare revenue. Companies House may have limited resources to chase up offenders.

▼ **TABLE 5:** AGE OF COMPANIES, MEAN AND MEDIAN YEARS OF ACTIVITY

	OTHER		DIGITAL ECONOMY	
	MEAN	MEDIAN	MEAN	MEDIAN
SIC 07	10.3	6.6	8.6	5.8
OF WHICH:				
ICT SECTOR			8.6	6.1
DIGITAL CONTENT			8.6	5.8
GI SECTOR AND PRODUCT	10.2	6.5	9.8	6.6

Note: Age defined as years of activity since the company was incorporated

Table 5 also provides results for the median company – that is, the one in the middle of the distribution. Looking at the median observation helps to highlight if the averages are influenced by outliers (in this case, very young or very old companies). We can see that the median differences are substantively smaller, which suggests outliers are present. Even so, the main result holds: on our GI-based measures, the median digital economy company is almost the same age as the median non-digital business.

We suspect there are two main reasons for these results. Firstly, when referring to the digital economy we should not only associate it with the rising importance of the ICT sectors but also with companies that are increasingly using digital technologies for their main activities (i.e. digital content industries). These companies are mainly working in the publishing sector, in the film and media production but also defence-space or aviation-aerospace sectors. They might have had a few computers in the office 30 years ago, but they would have been working mainly in a non-digital environment. They were able to adapt their production technologies to this change, in particular, the spread of ICT across workplaces, has helped broaden the scope of the digital economy.

The second reason is related to the structure of the dataset. We are comparing active companies, and have stripped out those who are in administration/liquidation, and it seems that digital economy status alone has no link to firm age.

CASE STUDY 2 - Kelton Engineering

London-based Stonewash started as a magazine publishing company but grew into a digital business. It now offers a platform for publishers to upload their offline content to mobile and desktop systems and provide the publisher extensive information about costumers. Their application integrates with social networks, expanding the relevant reader's information available to the publisher. Although operating at the core of the digital economy, the company's SIC Code 58190 – 'Other publishing activities' tells little about the company's main activity and competence. Growth Intelligence classifies Stonewash as a 'Custom Software Development' providing further insight about the company's business.

In Table 6, we show the distribution of companies by age groups. This share can easily be interpreted as a survival rate as nothing is revealed about the actual turnover rate of companies[9]. Panel A uses SIC code definitions of the digital economy, and provides the benchmark for our preferred definition (Panel B). In Panel B, around 63% of digital companies are under 10 years old, 30.5% under five years, 13.2% under three years old and around 0.9% less than a year old. This compares with 64.7%, 30.3%, 12.9% and 0.8% respectively in the rest of the economy. Analysing the distribution using SIC codes (Panel A) shows very similar patterns.

Start-ups are usually defined as companies less than three years old. If we look at the share of start-ups in the digital economy (as a share of all digital companies) and compare it with that of the rest of economy (as a share of all companies), we find that they are very similar results.

On the face of it, these findings are surprising. The popular image of the digital economy is of start-ups and very young companies, many of which are not durable bulwarks of the economy. This evidence also suggests that there is no reason to think that the digital companies are more ephemeral than the other companies. Our analysis of inflows, below, also tells a similar story. In the next phase of our research we hope to dig into this issue using other layers of Growth Intelligence's software, to help us identify companies that have moved from 'analogue' to digital platforms and production techniques.

▼ **TABLE 6:** DISTRIBUTION OF COMPANIES BY AGE GROUPS

	%	
	OTHER	DIGITAL ECONOMY
A. SIC 07		
UP TO 1 YEAR OLD	0.82	0.91
UP TO 3 YEARS	12.80	14.39
UP TO 5 YEARS	29.99	33.17
UP TO 10 YEARS	64.30	65.78
B. GROWTH INTELLIGENCE SECTOR AND PRODUCT		
UP TO 1 YEAR OLD	0.82	0.90
UP TO 3 YEARS	12.91	13.25
UP TO 5 YEARS	30.27	30.48
UP TO 10 YEARS	64.70	62.85

Note: Each entry represents the share of companies within each age group

[9] We have looked at companies that dissolved in year 2012, which have dropped from the selected sample. We have looked at the distribution of companies by incorporation year and by sector and also in this case, the distribution over time is similar in the digital sectors and in the rest of the economy. This also implies that the average age is similar and it is actually higher for the digital economy sectors when using Growth Intelligence definition.

▼ **FIGURE 1:** INFLOW OF COMPANIES BETWEEN 1980 AND 2011

4.2 INFLOWS

Figure 1 shows the inflow of companies into the economy, comparing inflows of companies in the digital economy (orange line) with companies in the rest of the economy (green line), from 1980 to 2012. The number of digital companies entering the economy every year has always been much smaller, but it is interesting to see that when using Growth Intelligence's classification we are able to capture a higher level of inflow starting from year 2000.[10] In Table A7, Panel A in the Appendix, we show the average flows over different periods. It has been increasing over time reaching its highest values in the period 2001 to 2011, for both digital and non-digital companies, according to any definition. We observe 12,888 digital companies (on average) entering the economy every year when using the conventional definition, as opposed to 18,926 when using our preferred Growth Intelligence's measure. [11]

We also estimate the growth rate, defined as the percentage of the yearly inflow over the total existing companies, and compare it across the two sectors. Results are shown in Figure 2 (the vertical line represents the burst of the dot-com bubble).

[10] Company reclassification may be more pronounced over longer periods: this will not be captured in SIC codes, which in Companies House are ascribed when companies are set up. Growth Intelligence's more up to date information may be buying us extra precision here.

[11] Four sectors ('information technology', 'architecture', 'marketing advertising', 'telecommunications') contribute to around 70% of the total flow of companies between 2001 and 2011. The patterns are very similar in all four sectors, but it is striking to see the important contribution given by the 'information technology' sector. Interestingly, architecture seems to be as important as the telecommunications in increasing the flow. However, it is important to bear in mind, that we are comparing sectors within our preferred definition of the digital economy.

Two things are worth noting. First of all, the growth rate of digital companies has been higher than the rate in the rest of the economy in the period before the dot-com bubble which happened in year 2000, and this is even more evident when using the SIC codes. The reason why the rate is smoother in Growth Intelligence's classification relates back to our earlier discussion: when using our alternative definition we are also capturing companies that have been in the economy for a longer period and have simply adapted to technological change. This can also be seen in Table A7 where between 1991 and 2000, we have a higher average growth rate of digital companies in the SIC codes measure (Panel B).

Second, after the burst of the dot-com bubble, the digital economy started to follow the cycles of the rest of the economy, and the growth rate even started to be lower than the rate in other sectors. In future versions of the research we will be able to explore sectoral turnover in more detail.

4.3 REVENUE

Regular Companies House data provides relatively limited information on company revenues.
Under Companies House rules, all public and private limited companies must file annual accounts and an annual tax return, which usually contains balance sheet information (see section A1 in the Appendix). However, private companies and business partnerships do not have to do this, and even among the companies obliged to report revenue information as part of a profit and loss account, not all do so.

Only 14% of the companies in our sample have reported revenues in the period between 2010 and 2012 and an even smaller percentage (8.6%) have filed revenues every year over the same period. This means there are substantial limits on what conventional analysis can do.

However, for this report we are able to use real-time modelled revenue information developed by Growth Intelligence, which covers all of the companies in the dataset.

▼ **FIGURE 2:** GROWTH RATE IN THE NUMBER OF COMPANIES 1980-2011

Note: Growth rate as a percentage of number of firms entering the economy each year over the total existing firms

▼ **TABLE 7:** MEAN AND MEDIAN REVENUES AND REVENUE GROWTH

	A. AVERAGE REVENUES						B. AVERAGE ANNUAL REVENUE GROWTH			
	COMPANIES HOUSE (%)		GROWTH INTELLIGENCE		OBS	SECTOR DISTRIBUTION (%)	COMPANIES HOUSE (%)		OBS	SECTOR DISTRIBUTION (%)
	MEAN	MEDIAN	MEAN	MEDIAN			MEAN	MEDIAN		
SIC 07										
OTHER	17,945,967	113,294	13,000,897	75,585	237,980	91.62	15.91	1.82	145,124	92.63
DIGITAL ECONOMY	11,496,924	105,634	11,295,655	98,261	21,771	8.38	19.90	3.58	11,545	7.37
GI SECTOR AND PRODUCT										
OTHER	18,380,097	110,048	13,218,321	71,677	227,430	87.57	15.68	1.70	138,750	88.56
DIGITAL ECONOMY	10,547,218	123,388	10,322,345	119,717	32,321	12.44	20.21	4.17	17,919	11.44

Note: Companies House average revenues are averaged over the period 2010 to 2012. Growth Intelligence revenues are computed over the same sample. For the Companies House dataset if for each company there is more than one observation, only the most recent is kept. Average annual revenue growth is computed on a smaller sample, as information for at least two consecutive years is need. The years considered are the same as above, 2010 to 2012.

To see how these estimates compare, Table 7 sets out the revenues for the companies that report revenues to Companies House, with the modelled revenues from Growth Intelligence. Panel A provides some basic descriptives. We can see that the sub-sample is similar to the full sample in terms of digital economy shares. This is shown in the last column of Panel A, where we report the sectoral distribution of the sample of companies that report revenues to Companies House. Using either SIC codes or Growth Intelligence classification, non-digital companies have higher average revenues, but on Growth Intelligence's measures the digital-non-digital gap narrows substantially for the average company. Interestingly, when shifting to the median digital economy company, observed and modelled revenues are higher than revenues for the median non-digital company (if using GI-based definitions to define the economy). SIC-based definitions give mixed results, depending on whether Companies House or Growth Intelligence revenue figures are deployed.

In Panel B of Table 7, we look at revenue growth for the sub-sample of companies who report revenues to Companies House over more than one year, during the period 2010-2012. The first column reports the average percentage growth, defined as the within-firm growth of revenues averaged over the sample. The rate is higher for digital companies (20%) against a rate of 16% in the rest of the economy. Similar results are obtained on the SIC benchmark. The same basic result holds when we look at the median company in the distribution, but the digital/non-digital differences in revenue change are a lot smaller.

On the face of it, our analysis suggests that while digital economy companies have lower average revenues than the rest of the economy, they exhibit higher median revenues, and higher average and median revenue growth. As we sketched out earlier, this result has to be treated with some care, since the modeled data is standing in for observed revenue among the many companies who do not report revenue to Companies House. We should also remember that this is not a random sample – for most companies, having some revenue to report is conditional on strong underlying performance. So there might be some positive selection at work, and the result does not imply a causal relationship from being in the digital economy to revenue levels or changes.

▼ **TABLE 8:** GROWTH INTELLIGENCE MODELLED REVENUES

GI (MEAN) REVENUES				
	SIC 07		GI SECTOR AND PRODUCT	
	MEAN	MEDIAN	MEAN	MEDIAN
OTHER	2,856,186	56,476	2,898,372	53,487
DIGITAL ECONOMY	1,824,130	60,362	1,888,099	77,105
OF WHICH:				
ICT SECTOR	2,773,946	68,586		
DIGITAL CONTENT	1,316,069	55,102		

Next, we look at modelled revenues across the full sample of companies – not just the sub-sample above. We use Growth Intelligence's modelled figures to do this. Table 8 shows the main results. As with the sub-sample analysis above, we can see that average revenues for all digital economy companies are running at about 60% of the average for all other companies. As before, when we measure the digital economy using the Growth Intelligence sector and product typology, the revenue gap grows slightly, but so does the average digital company's revenue. The increase in average revenues is due to the increase of companies in sectors like 'computer games', ' animation' or 'consumer electronics', with product types like 'software web application', 'software desktop or server', 'web hosting' and 'custom software development'. However, when we look at the median firm, the basic result reverses, as it does for the sub-sample: median modeled revenues are higher for digital economy firms than non-digital, with a bigger gap when GI-digital definitions are used. The median firm's modeled revenue is always a great deal smaller than the average, suggesting a few very high-revenue firms are pulling the means up.

4.4 EMPLOYMENT

Under Companies House rules, companies are only obliged to report employment data in specific cases (e.g. SMEs needing to prove their status). Employment data is not required in annual returns or accounts. Even more than with revenue data, this places obvious limits on what researchers can do to look at registered companies' employment and job creation activities.[12]

As with companies reporting revenue, this is not a completely random sub-sample; having employees is conditional on being big enough to pay staff; but also not all companies with employees report this information. We therefore perform some diagnostic checks to compare our sub-sample with the bigger sample of companies. We would expect companies with employees to be older and have higher revenues than those without, and this turns out to be the case: those in the employment 'set' are on average twice as old, and report average modelled revenues around 2/3 higher than the non-employment 'set'. But, tests of industrial structure suggest very similar shares of digital and non-digital companies and the spatial distribution of the companies across the UK is very similar, with three out of the top five locations being shared.[13] So, although the employment sample is biased towards older and larger companies, it is not clear that this necessarily results in differential bias across digital and non-digital companies. Even so, what follows should be treated with some care.

[12] We experiment with boosting the sample by adding in employment information from social media sources. This increases coverage to over 143,000 companies. However, in its current form this data only has approximate date information, and resulting employment estimates are unstable. We hope to improve on this in future versions of the research.

[13] Full results are available on request.

▼ **TABLE 9:** AVERAGE EMPLOYEES PER FIRM

	BREAK DOWN	GI SECTOR* PRODUCT		SIC CODES		OBS
		MEAN	MEDIAN	MEAN	MEDIAN	
2008-2012	OTHER	28.19	5	28.66	5	134,767
	DIGITAL ECONOMY	28.33	5	21.11	4	
2010-2012	OTHER	20.40	4	20.94	4	70,723
	DIGITAL ECONOMY	23.37	4	17.23	3	

Note: Sub-sample of companies filing employment information to Companies House

▼ **TABLE 10:** DIGITAL ECONOMY EMPLOYMENT SHARES

CATEGORY	SHARE OF ALL EMPLOYMENT (%)	
	2008-2012	2010-2012
DIGITAL ECONOMY (SIC CODES)	5.10	5.08
OTHER	94.90	94.92
DIGITAL ECONOMY (GI PRODUCT*SECTORS)	11.39	11.33
OTHER	88.61	88.67

Notes: Sub-sample of companies filing employment information to Companies House

First, we look at employees per firm. Table 9 shows average employment per company and median employment per company, inside and outside the digital economy. As not all companies report employment in every year, the data is smoothed across a five-year and a three-year period. We can see that average employment counts differ substantially depending on how the digital economy is measured. Using conventional SIC measures, average employee counts are around a third smaller for digital economy companies than non-digital. By contrast, using our preferred Growth Intelligence-based definitions, digital economy companies employ, if anything, slightly more people than companies in the wider economy.

Looking at median employment per company, though, there is no difference between digital and non-digital companies using Growth Intelligence definitions (although there are slight differences using SICs). Medians represent companies in the middle of our sample, and so this suggests there are some outliers in both digital and non-digital sectors explaining some of the mean differences.

Second, we turn to the digital economy's share of all employment (for which we have information). In this case the numbers are broadly the same (Table 10), and the main message is identical: shifting from SIC-based measures of the digital economy to Growth Intelligence's measures shifts the digital economy's employment shares substantially upwards, from around 5% to around 11%. This is as we would expect, since this also increases the number of companies.

We also experiment with different estimates of employment growth, drawing on both observed information (from Companies House) and modelled information on recent/current/projected employment change. The latter is developed by Growth Intelligence from transactional data related to company purchasing behaviour during the sample period, combined with information on company characteristics and machine-learnt benchmarks, to estimate the likely change in employees. In both cases, the sub-sample of companies is currently small, and results are not yet stable. We hope to develop this line of enquiry further in future work.

05 WHERE IS THE DIGITAL ECONOMY?

This section looks at the spatial footprint of the digital economy.

5.1 FIRM COUNTS BY TTWA

To get a sense of how the digital economy is distributed across the UK, we geo-code individual companies into Travel to Work Areas (TTWAs). TTWAs are designed to represent functional labour markets, and are generally considered to the best available approximation of a local economy.[14] We also do this because we have the registered addresses for companies; in most cases this will also be their trading address, but not all. Bundling companies into TTWAs minimises the chances of putting companies in the 'wrong' part of the country (although we can't rule this out in a few cases)[15].

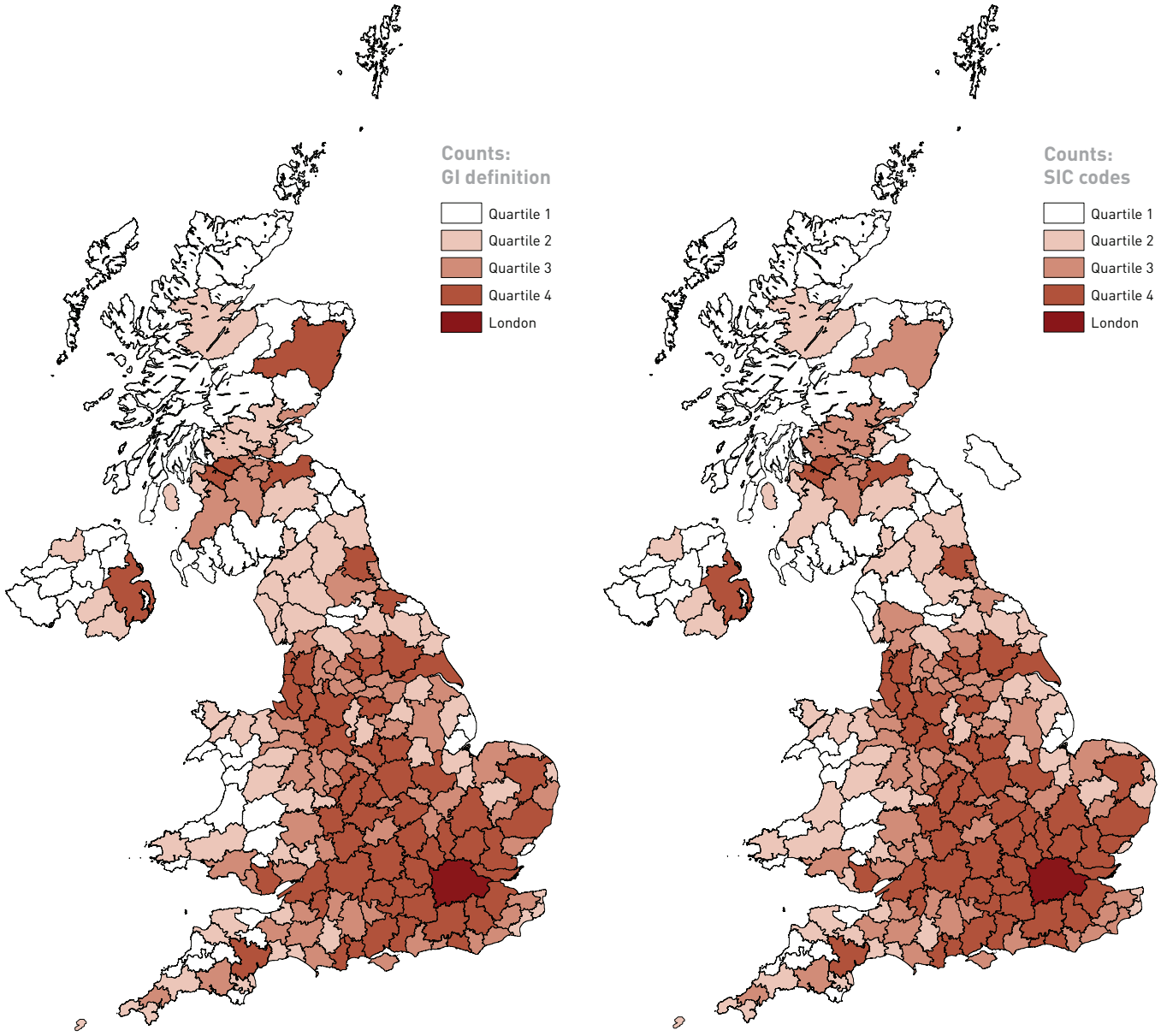
Geo-coding slightly shrinks our benchmarking sample, from 1.868m to 1.771m companies. This is because not all company addresses provided to Companies House include postcodes, and because some companies provide PO Box addresses (where the postcodes are not assigned to a particular geography).

We first look at the distribution of companies around the country. Figure 3 maps the UK's Travel to Work Areas and shows banded counts of digital economy firms. The left hand panel shows the results when we use Growth Intelligence-based measures; the right hand panel provides a benchmark distribution using SIC-based measures. For easy comparison we have divided the counts into quartiles, each of which represents 25% of the observations, plus a separate London band.

The digital economy is very spiky, with a lot of co-location in London, Manchester and the Greater South East. Using our preferred Growth Intelligence sector*product measure, the 10 TTWAs with the most digital economy companies are London (64,630 companies), Manchester (7,324), Guildford and Aldershot (6,158), Luton and Watford (5,147), Wycombe and Slough (4,979), Birmingham (4,695), Reading and Bracknell

[14] Formally, 75% of those living in a given TTWA also work in the TTWA, and vice versa.
[15] As noted above, using registered companies will understate the numbers of business units 'on the ground'. For area analysis, also likely to under-count the number of registered companies in smaller places which are less likely to have head offices and boost numbers in bigger cities, especially London. In future versions of the research we will be able to deal with these issues by working with trading address information, as well as companies' registered addresses.

▼ **FIGURE 3:** COUNTS OF DIGITAL ECONOMY COMPANIES BY TRAVEL TO WORK AREA, 2012.

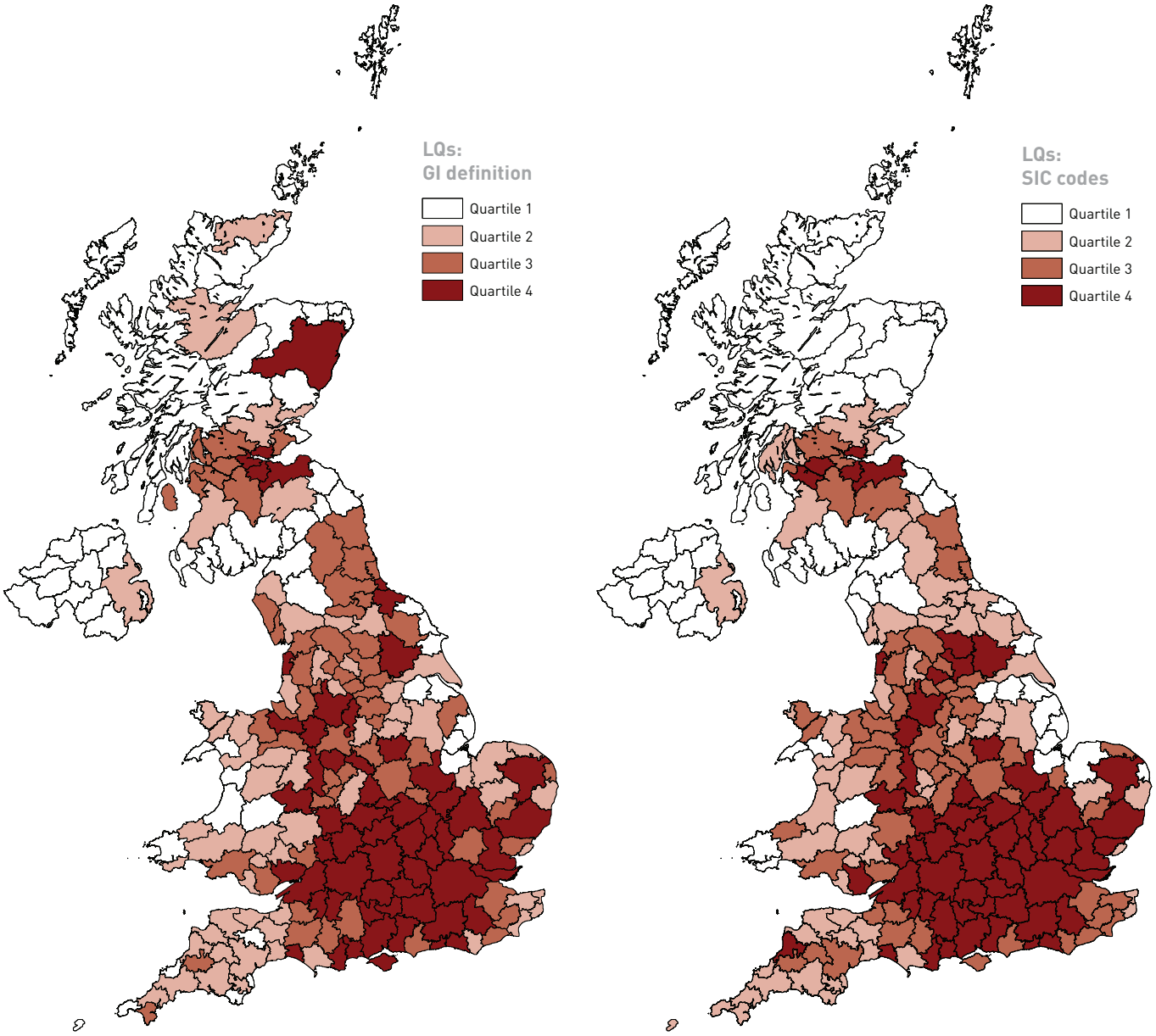


(4,914), Bristol (4,714), Crawley (3,867) and Brighton (3,730). Underneath this group is another 50 TTWAs with over 1,000 digital economy companies, followed by a very long tail: over half the areas on the map have less than 600 companies, and 25% have under 200.

Using digital economy SIC codes the top 10 TTWAs are very similar: London (51,491 companies), Manchester (4,737), Guildford and Aldershot (4,489), Luton and Watford (3,908), Reading and Bracknell (3,823), Wycombe and Slough (3,648), Bristol (3,233), Birmingham (3116), Brighton (2,992) and Crawley (2,751).

Overall, around 80% of companies are in urban areas - defined as TTWAs with a city of at least 125,000 people. Remember that in both cases, these are counts from our benchmarking sample.

▼ **FIGURE 4:** LOCATION QUOTIENTS OF DIGITAL ECONOMY COMPANIES BY TRAVEL TO WORK AREA, 2012.



5.2 CLUSTERING ACROSS THE UK

Next, we use location quotients to get a sense of where the digital economy is most locally clustered (in the sense of co-location). Location quotients compare the local area share of a group i to its national share. [16] Location quotients over 1 indicate local clustering; under 1 de-clustering. As before, the left hand map shows Growth Intelligence-based results; the right panel gives a benchmark using the SIC-based definitions.

Looked at this way, the spatial footprint of the digital economy is rather different. Using our preferred Growth Intelligence-based metrics, the 10 areas with the highest location quotients are Basingstoke

[16] Formally, LQia = (pia / pa) / (pi / p), where pia / pa is the local population share of i in area a, and pi / p is i's national population share. An LQ of above 1 indicates concentration, or local shares above the national shares; scores below 1 indicate dispersion, or local shares below the national share.

(1.66), Reading (1.54), Newbury (1.52), Milton Keynes and Aylesbury (1.42), Poole (1.41), Stevenage (1.343), Guildford and Aldershot (1.36), Luton and Watford (1.34), Wycombe and Slough (1.33), Bournemouth (1.32). Brighton also scores 1.32, with Cambridge just behind on 1.30. Using LQs, then, highlights the importance of the digital economy to cities in the Greater South East, especially in the crescent of high-value activity that runs around the West of London.

SIC and Growth Intelligence-based measures also differ. Using SIC codes to define the top 10 digital economy clusters we find a broadly similar pattern: the 10 most digitally-clustered TTWAs are Basingstoke (1.77), Reading (1.73), Newbury (1.60), Milton Keynes (1.529), Brighton (1.528), Luton and Watford (1.47), Guildford (1.44), Wycombe and Slough (1.40), Stevenage (1.39) and Andover (1.36). However, the map shows a rather different picture for the UK as a whole. Defining the digital economy using only SIC codes turns pretty much the whole of Southern England into a digital hotspot, and obscures the real extent of co-location in areas like Aberdeen and Middlesbrough.

Why don't we find places like London, Manchester and Birmingham in these lists too? Partly because these are large cities with diverse economies; there's a lot going on besides digital economy activity. Also because digital economy companies often cluster at much smaller scales within cities - for example, in Silicon Roundabout (Nathan, Vandore et al. 2012). We will explore these very local hotspots in more detail in future versions of the research.

06 DISCUSSION

Our main findings are:

- The digital economy is poorly served by conventional definitions and datasets. Big data methods can provide richer, more informative and more up to date analysis.

- Using Growth Intelligence data on a benchmarking sample, we find that the digital economy is substantially larger than conventional estimates suggest. On our preferred measure, it comprises 14.4% of active companies in the UK (as of August 2012). This compares to 10.0% of companies when conventional SIC-based definitions are used.

- SIC definitions of the digital economy miss out a large number of companies in business and domestic software, architectural activities, engineering, and engineering-related scientific and technical consulting, among other sectors.

- Companies in the digital economy have a similar average age to those outside it. Shares of start-ups (companies up to three years old) are very similar. Given the popular image of the digital economy as start-up dominated, this may be surprising to some. The result partly reflects wider technological shifts: as digital platforms and tools spread out into the wider economy, and become pervasive in a greater number of sectors, so the set of 'digital' companies widens.

- Inflows of digital companies into the economy have always been relatively small, given its sectoral share. However, using our new definitions of the digital economy, inflow levels are substantially higher.

- As far as we can tell, digital economy companies have lower average revenues than the rest of the economy, but the median digital company has higher revenues than the median company elsewhere in the economy. Revenue growth rates are also higher for digital companies. However, these results come from a sub-sample of older, likely stronger-performing companies, so there is some positive selection at work.

- Switching from SIC-based to Growth Intelligence-derived measures substantially increases the digital economy's share of employment, from around 5% to 11% of jobs. Digital economy companies also show higher average employment than companies in the rest of the economy (this reverses when we use conventional SIC-based measures of the digital economy). Looking at median employees per firm, though, the digital/non-digital differences are always a lot smaller. Our employment results should also be treated with some care, as not all companies report their workforce information.

- The digital economy is highly concentrated in a few locations around the UK: Growth Intelligence software provides a fresh look at these patterns. In terms of raw firm counts, London dominates the pictures, but Manchester, Birmingham, Brighton and locations in the Greater South East (such as Reading and Crawley) also feature in the top 10. Location quotients show the extent of local clustering, which for the UK's digital economy is highest for areas in the Western arc around London, such as Basingstoke, Newbury and Milton Keynes. But areas like Aberdeen and Middlesbrough also show high concentrations of digital economy activity.

There are a number of issues we do not look at here, and which we hope to address in future research. These include:

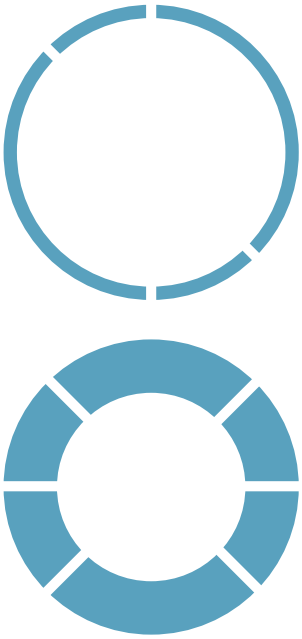
- Productivity of companies in the digital economy vs. other sectors;

- Turnover of companies (i.e. net births) in the digital economy vs. other sectors;

- Multiplier / spillover effects of the digital economy on business creation, employment and revenues in other sectors;

- Potential displacement effects of the digital economy on business creation, employment and revenues in other sectors.

07 REFERENCES & APPENDICES



REFERENCES

Aumeyr, M. (2008). Data Linkage: Financial Analysis Made Easy and Workplace Employment Relations Survey. Newport, ONS.

Cabinet Office (2012). Unlocking growth in Cities - City Deals: Wave 1. London, Cabinet Office.

Cable, V. (2012). Industrial Strategy. Speech to Imperial College London. London, 11 September.

Cameron, D. (2010). East End Tech City speech, www.number10.gov.uk/news/east-end-tech-city-speech/ (accessed 29 May 2012).

Centre for International Economics (2005). Australian digital content industry futures. Report prepared for the Department of Communications, Information Technology and the Arts (DCITA). Canberra, CIE.

Choi, H. and H. Varian (2012). "Predicting the Present with Google Trends." Economic Record 88: 2-9.

Clark, G. (2012). City Deals Wave Two Launch speech. London, 29 October.

Department for Business Innovation and Skills (2012a). Business Population Estimates for the UK and Regions, 2012: Methodology and Quality Note. BIS Enterprise Directorate. London, BIS

Department for Business Innovation and Skills (2012b). Business Population Estimates for the UK and Regions 2012. London, BIS.

Department for Business Innovation and Skills (2012c). Industrial Strategy: UK sector analysis. London, BIS.

Department for Business Innovation and Skills and Department for Culture Media and Sport (2009). Digital Britain: Final Report. London, BIS.

Department for Business Innovation and Skills, Department for Culture Media and Sport, et al. (2010). Impact Assessment for the Digital Economy Bill. London, BIS.

Einav, L. and J. D. Levin (2013). The Data Revolution and Economic Analysis. National Bureau of Economic Research Working Paper Series No. 19035. Cambridge, MA, NBER.

Foord, J. (2013). "The new boomtown? Creative city to Tech City in east London." Cities (Online First).

King, G. (2013). Restructuring the Social Sciences: Reflections from Harvard's IQSS. Cambridge, Mass, Institute for Quantitative Social Science.

Lucas, R. (1988). "On the Mechanics of Economic Growth." Journal of Monetary Economics 22: 3-42.

Moretti, E. (2012). The New Geography of Jobs. Boston, Haughton Mifflin Harcourt.

Nathan, M. (2011). "East London Tech City: Ideas without a Strategy." Local Economy 26(3): 197-202.

Nathan, M., E. Vandore, et al. (2012). A Tale of Tech City: The future of East London's digital economy. London, Centre for London.

Office of National Statistics (2007). UK Standard Industrial Classification of Economic Activities 2007 (SIC 2007): Structure and explanatory notes. Newport, ONS.

Office of National Statistics (2010). Business Structure Database: User Guide. Newport, Social and Economic Micro Analysis Reporting Division, Office for National Statistics.

Office of National Statistics (2012). Guide to the Business Population and Demographic Statistics Publications. Newport, ONS.

Romer, P. (1990). "Endogenous Technological Change." Journal of Political Economy 98(5): 71-102.

A1/ ABOUT COMPANIES HOUSE DATA

Under the Companies Act 2006, all limited companies in the UK, and overseas companies with a branch or place of business in the UK need to be registered with Companies House. Some business partnerships (such as Limited Liability Partnerships) also need to register. There is a charge of around £100 to do this. Sole traders and business partnerships which are not LLPs do not need to register at Companies House, although they will need to file tax returns with HMRC. When they register, companies are asked to choose the Standard Industrial Classification (SIC) code which best reflects their principal business activity. Dormant and non-trading companies are also asked to include SIC information.

All registered companies must file a) annual company returns as well as b) annual financial statements (statutory accounts).

Returns cover details of directors and company secretary, registered office address, shares and shareholders, as well as company type and principal business activity. There is a small charge for filing the return, which must be done within 28 days of the anniversary of incorporation. There are financial penalties for not filing the return on time: in the extreme Companies House can dissolve the company and prosecute the directors.

Statutory accounts must be filed with Companies House, in addition to tax returns with HMRC. Accounts must include a balance sheet, a profit and loss account, a directors' report and an auditors' report. The balance sheet shows the value of company assets; the profit and loss accounts shows sales, running costs and subsequent profit / loss. Accounts must be compiled by nine months after the end of the financial year. As with returns, there are financial penalties for late filing, and possible criminal penalties for non-filing.

A number of companies are exempted from full filing. Limited companies that are 'small' can send abbreviated accounts consisting only of the balance sheet, and in some cases can apply for exemption from auditing. Small firms must meet two or more of the following: less than £6.5m turnover; less than £3.26m on the balance sheet; fewer than 50 employees. Some 'dormant' limited companies can also claim partial or full exemption from filing. Dormant companies are those defined as having no 'significant accounting transactions' during the accounting period in question.

Companies must inform Companies House about changes to limited companies, including directors / secretaries joining or leaving; changes to the company name, registered address or accounting dates, and where records are kept. Limited companies can request to be closed / dissolved, providing they have not traded within the last three months; not changed company name within that period; are not subject to current / proposed legal proceedings, and have not made a disposal for value of property or rights. There is a £10 charge for the striking off application. Once Companies House has accepted the application, a notice is placed in the London / Edinburgh / Belfast Gazette giving at least three months' notice of the intent to remove the company from the Register.

A2/ THE GROWTH INTELLIGENCE DATASET

The Growth Intelligence dataset was created by software developed by Growth Intelligence, a company based in London. Growth Intelligence tracks the performance and activity of businesses using novel data sources and methods. Previous ways of monitoring companies have relied on companies to submit their own accounts and classifications. By contrast, Growth Intelligence detects the digital 'signature' that companies emit and forms conclusions based on the simultaneous observation of millions of businesses in real-time.

A2.1 Data-gathering and modelling

Growth Intelligence gathers 'document sets' for every company in the economy, for example website text, company news etc. Growth Intelligence then uses a suite of techniques known as text mining to classify and tag companies according to their own classification system. Just as document sets change every week as new products are launched and companies change what they do, how they think of themselves and how other people think of them, Growth Intelligence re-classifies companies continuously to make sure information is up-to-date.

Growth Intelligence uses techniques commonly referred to as 'predicting the present' or 'nowcasting' to provide intelligence on the performance of companies in real-time. Companies House provides historical information on company performance as reported by a company's accountants. However, this information is on average 16 months old. Companies reveal much richer information about themselves, their performance and their activity, directly and indirectly online. Growth Intelligence software uses bespoke sensor networks, structured web crawling and connections to existing data sources to track company performance in real-time. Examples of these non-traditional and semi-structured 'big' data sources include email traffic, social media activity, web traffic, search engine traffic, financial transactions and the semi-structured reporting of train and lorry spotters to identify freight and the activity of transport networks (the so-called 'Internet of Anoraks').

Growth Intelligence uses bespoke machine learning to learn the correlations between activity, real-time classification, digital signature and performance. Different signals are important for different types of business. For example, for offshore wind farms, a single metric – the amount of land they are granted by the Crown is – a very good predictor of revenue. For most other businesses this signal is useless. Because digital signatures change constantly, Growth Intelligence's imputed revenue figures also change in real-time. For the user of Growth Intelligence this translates into a 'nowcasted' growth curve (high-growth, growth, neutral growth or negative growth) right up to the present day. For this report, data for a single day was selected.

Each company's resulting 'digital signature' is composed of thousands of independent signals. Because very few companies have absolutely no digital signature (just as very few people are perfectly 'off the grid'), the number of 'invisible' companies is negligible.

Because Growth Intelligence classifications and growth figures are derived using automatic tracking of a large number of independence sources and machine learning models based on millions of observations, all companies are analysed using a single system and the results are therefore more complete and objective.

A2.2 Benefits of the approach

First, companies are all analysed in the same objective way, and there is much fuller coverage of company types than is available in SIC codes. SIC codes are designed to represent a firm's 'principal business activity' - the activity that contributes most to value added (ONS, 2007). SICs thus primarily represent information about outputs ('the character of goods and services produced'), but also aggregate information about inputs ('inputs, processes and technologies'), and clients ('the uses to which these are put'). Ideally, statisticians would allocate a SIC code based on detailed information about organisations' business models and their contribution to value-added.

However, in Companies House data, SIC code datasets rely on a company director or their accountant taking the time to select the most relevant numerical code from hundreds of possibilities. This means that every company is classified by a different person which their own interpretation of the code. There is very little overt incentive to classify a company 'correctly' and there is no practical way to detect and penalise 'wrong' classifications. The SIC system is also outdated, with much more detail in secondary industries than tertiary. As a result, over 30% of businesses have either no reported SIC code (21%), or an uninformative code (9.2% only considering: Other business support service activities not elsewhere classified, Other service activities not elsewhere classified or Other Business Activities). A further group will have assigned themselves incorrect SICs given their business fundamentals.

Second, some of the key modelled variables may be more robust than observed information. Businesses can shift Companies House reported turnover to suit immediate needs (for example to avoid competitors knowing how they are performing) using normal and legal accounting practices. Using a range of thousands of signals – many very hard to manipulate – makes it harder for a company to manipulate their revenue figures in Growth Intelligence.

Third, data is kept up to date. Document sets change every week as new products are launched and companies change what they do, how they think of themselves and how other people think of them; Growth Intelligence re-classifies companies continuously to make sure it is up-to-date.

A3/ USING THE GROWTH INTELLIGENCE DATASET

This section deals with data cleaning and the construction of our 'benchmarking' sample; then sets out the steps we take to define the digital economy using the Growth Intelligence dataset.

A3.1 Constructing a benchmarking sample

We start with 3.07m registered companies in the Growth Intelligence dataset as of August 2012. We have two aims in our sample-building. First, we need to clean the data so that it roughly corresponds to the relevant business population - removing, for example, dormant, dissolved and shell companies and removing duplicates (some firms may set up a number of registered companies - for example, a retail and a wholesale operation - we treat these as a single firm).

Second, we need to create a sample that allows benchmarking between SIC-based definitions of the digital economy and alternative industry definitions (hence this is a 'benchmarking sample'). Note that because around 21.5% of companies (661,859 observations) in the underlying Companies House data do not provide SIC information, the benchmarking sample will be smaller than the 'true' number of companies in the UK. Below we set out some basic diagnostic checks using the Business Structure Database, to see what would result from our cleaning steps without the benchmarking requirement.

Our cleaning and benchmarking steps are as follows:

Step 1: we drop all companies who are non-trading and those who are dormant, with no significant trading activity for the past 12 months. Using information from the London and Edinburgh Gazettes, we also drop dissolved companies, companies in any kind of receivership or administration. We keep companies that are active but in the process of striking off. The reason for keeping these firms in is mainly that most of them are still operative in the market. Secondly, some of these companies may be in this status because they failed to file their returns to Companies House but may re-emerge in the market under a different name.

Overall, this reduces our sample to about 2.556m companies. When these steps are run separately, dropping non-trading companies removes 92,929 observations; dropping dormant companies removes 106,589 observations; dropping all but active and partially active companies removes 318,906 observations.[17]

Step 2: we drop holding companies from the sample, as these by definition simply control other companies. This reduces our sample to 2.546m observations.

Step 3: we identify related companies, and keep only the largest component of the group. We do this in order to avoid multiple counting in the underlying firm structure; and to avoid multiple counting in key measures such as revenue and employment (for instance, if company A controls company B, it may include some of B's revenue / employment in its accounts). Specifically, we group companies on the basis of name (same name)[18], postcode of registered address (same location) and SIC5 code (same industry). Within each group thus identified, we keep the unit reporting the highest revenue (as modelled by Growth Intelligence). Note that for the purposes of benchmarking, we are required to do the industry matching on SIC code. This procedure gives us a substantially smaller set of companies (reducing the benchmarking sample to 1.868m observations), and will also give us conservative estimates of revenue and employment.

We cross-check the reliability of our chosen sampling method. Specifically, we use a less fine geographic detail found in the postcode including only information up to the sector level (i.e. the first 4/5 letters). The reason we are checking on this is related to the fact that there are some companies that have the same name, same SIC code, but for some reason are not located at exactly the same postcode unit (there are usually 15 addresses per unit). Therefore they would be part of two different groups. Yet, in the less restrictive approach if they have the same 4/5 digit of the postcode, they would be included in the same group. On the other hand, the drawback of this approach is that we might include companies that operate in the same industry, are located in the same broader area, have the same first (second or third) word in their name, but are not in practice in the same group.

[17] Some companies may be in more than one of these categories, so sub-totals may not sum.

[18] We use the first word if it is not a single letter, otherwise the second and up to the third and we exclude acronyms and abbreviations that refer to the status of the company (Limited, Ltd, Plc, Company, LLP) in all their forms.

Results from this alternative procedure show that the number of companies decline but almost proportionally in both the digital economy and in the rest of the economy (results are available on request). This is reassuring, as it implies that there is nothing systematic happening in the selection process we are using that under/overcounts the companies differently in the digital economy and in the rest of the economy.[19]

Finally, we briefly discuss two wider issues here. First, what are the effects of grouping on SIC codes? If the process generating missing codes is random and therefore not related to any characteristics of the company, we are still confident that our sample produces reliable estimates of the digital sector, yet if it is non-random we are likely to face a more important issue, as it could have different impact on the counting of companies in different sectors. In the latter case, the fraction of companies in one or the other sector would be biased. If, for example, companies that act within the digital economy were the ones more likely not to report a SIC code (they cannot find any code to describe their activity), we would be underestimating the number of companies in the digital economy, but not the number of companies in the other sectors.

Second, how much smaller is our benchmarking sample than the 'true' sample of companies, and how accurate is the latter? Recall that over 660,000 of 3.07m raw observations have no SIC information. Repeating the cleaning exercise above using Growth Intelligence sector categories to help identify groups - rather tha SIC codes - produces a sample of 2.254m companies.

To get a sense of whether this 'true' sample is accurate, we compare our numbers against estimates for the complete set of registered enterprises in the UK, drawn from the Business Population Estimates (BPE) published by BIS (Department for Business Innovation and Skills 2012b). The BPE offers a useful benchmark, for two reasons. First, it combines information from the IDBR/BSD (VAT and PAYE-registered firms) with an estimate for the set of unregistered businesses in the UK, drawn from HMRC tax returns and the Labour Force Survey. This latter group will include a number of companies and partnerships that file at Companies House but do not appear in the IDBR because they fall below revenue and/or employment thresholds. Second, company-level information is also transformed into enterprise-level, using cleaning steps to simplify corporate structures (Department for Business 2012a; Office of National Statistics 2012).[20]

BPE data is currently available from 2000-2012; the latest figures estimate 1.8m 'companies and partnerships' in the UK as of the start of 2012, plus 3.1m 'sole proprietorships and partnerships' comprising 0.3m with employees, and 2.71m consisting of only the self-employed owner-manager, or only an employee director.

To determine our benchmark figure, we proceed as follows. We start with the 1.8m 'companies and partnerships', as these enterprises will all file with Companies House. The 'sole proprietorships and partnerships' group will also include some enterprises who appear in Companies House, so we will need to include some of them. We take a conservative approach and include only those with employees. This gives us 2.1m enterprises at the start of 2012.

[19] Alternatively, grouping could be performed using shareholder information. The underlying intuition is similar: if company A owns more than (say) 50% of company B, only A is kept in the sample. We plan to run a sub-sample robustness check using shareholder information in future versions of the research.

[20] The ONS reports that only around 26% of enterprises can be directly matched to a single Companies House ID; more than a third of enterprises match to more than one ID; 8% of enterprises match to 20 or more IDs (Aumeyr 2008).

▼ **TABLE A1:** OFFICIAL DIGITAL ECONOMY SIC CODES

SIC2007CODE	DESCRIPTION
ICT SECTOR	
28230	MANUFACTURE OF OFFICE MACHINERY AND COMPUTERS
26200	MANUFACTURE OF COMPUTERS AND OTHER INFORMATION PROCESSING EQUIPMENT
27320	INSULATED WIRE AND CABLE
26110	ELECTRONIC VALVES AND TUBES AND OTHER ELECTRONIC COMPONENTS
33200	TELEVISION, RADIO TRANSMITTERS AND APPARATUS FOR TELEPHONY AND TELEGRAPHY
26400	TELEVISION AND RADIO RECEIVERS, SOUND OR VIDEO RECORDING OR PRODUCING APPARATUS AND ASSOCIATED GOODS
26511	INSTRUMENTS AND APPLIANCES FOR MEASURING, CHECKING, TESTING AND NAVIGATING AND OTHER PURPOSES
26512	INDUSTRIAL PROCESS EQUIPMENT
46439	WHOLESALE OF ELECTRICAL HOUSEHOLD APPLIANCES
46510	WHOLESALE OF COMPUTERS, COMPUTER PERIPHERAL EQUIPMENT AND SOFTWARE
46660	WHOLESALE OF OTHER OFFICE MACHINERY AND EQUIPMENT
46520	WHOLESALE OF OTHER ELECTRONIC PARTS AND EQUIPMENT
46690	WHOLESALE OF OTHER MACHINERY FOR USE IN INDUSTRY, TRADE AND NAVIGATION
61900	TELECOMMUNICATIONS SERVICES
77330	RENTING OF OFFICE MACHINERY AND EQUIPMENT INCLUDING COMPUTERS
62020	COMPUTER HARDWARE CONSULTANCY
95110	MAINTENANCE AND REPAIR OF OFFICE, ACCOUNTING AND COMPUTING MACHINERY
62090	OTHER COMPUTER RELATED ACTIVITIES
DIGITAL CONTENT INDUSTRIES	
58110	PUBLISHING OF BOOKS
58130	PUBLISHING OF NEWSPAPERS
58142	PUBLISHING OF JOURNALS AND PERIODICALS
59200	PUBLISHING OF SOUND RECORDINGS
58190	OTHER PUBLISHING
18110	PRINTING OF NEWSPAPERS
18129	PRINTING N.E.C
18130	PRE-PRESS ACTIVITIES
18130	ANCILLARY ACTIVITIES RELATING TO PRINTING
18201	REPRODUCTION OF SOUND RECORDING
18202	REPRODUCTION OF VIDEO RECORDING
18203	REPRODUCTION OF COMPUTER MEDIA
58290	PUBLISHING OF SOFTWARE
62020	OTHER SOFTWARE CONSULTANCY AND SUPPLY
63110	DATA PROCESSING
63110	DATABASE ACTIVITIES
73110	ADVERTISING
74209	PHOTOGRAPHIC ACTIVITIES
59111	MOTION PICTURE AND VIDEO PRODUCTION
59131	MOTION PICTURE AND VIDEO DISTRIBUTION
59140	MOTION PICTURE PROJECTION
59113	RADIO & TV (DCMS ESTIMATES)
63910	NEWS AGENCY ACTIVITIES

Source: BIS (2009)

To estimate the number of enterprises in August 2012, we calculate the change in enterprise numbers between 2011-2012 (six percentage points) and assume this also holds in 2012. We smooth the growth through the year, so for August, the projected increase will be (0.06/12)*8 = 4 percentage points on the January figure. This gives us a projected count of 2.184m enterprises in August 2012.[21]

This figure is slightly lower than our 'true sample' of 2.254m companies, but within 96.9% of the figure. In turn, this suggests our cleaning steps give broadly accurate results.

A3.2 Redefining the digital economy

The 'official' definition of the digital economy is done using SIC codes, as reported in Table A1. As we set out in the main body of the report, there are substantial challenges in using SIC codes this way, and good reason to think resulting estimates will contain error.

We are able to use an alternative, multi-dimensional classification system developed by Growth Intelligence. Growth Intelligence classifies companies by their 'sector context', 'product type', 'sales process' and 'client type' (amongst others). These codings are probabilistic, not 0/1 as in SIC codes, and the 'most likely' categories are used here.

A major advantage of the Growth Intelligence system is thus that it allows us to look inside overall sector 'boxes'. Specifically, we can describe companies by what they sell (product type), how they sell (sales process), to whom they sell (client type) and in the context of what value chain they sell (value chain context). We exploit this feature in our definition strategy. [22]

In essence, we develop a 'rough cut' of the digital economy using conventional SIC codes, look at the corresponding Growth Intelligence sector and product groups for those firms, then map the sector-by-product categories back onto our full benchmarking sample. The aim is to identify companies in digital sector contexts whose main outputs (products and services) are also digital. For example, we want to exclude large parts of the architecture sector, but include firms specialising in CAD and technical drawing. By the same token, we want to exclude supermarkets, but include retailers whose principal offering is digital (such as digital music stores). In this way we exploit the richness of the Growth Intelligence information (using sectors and products, not just sectors), while cleaning out the error in the SIC-based classifications.

We build our new definition of the digital economy in six stages.

Step one: We take the sample of companies who have digital economy SIC codes, and map each company identified by the conventional definition onto its corresponding Growth Intelligence sector and product.

We need to have some initial list of industrial sectors that refer to the digital economy, and use the Government's definition of the digital economy as our starting point - as it bears some resemblance to what most commentators would consider to be 'the digital economy'. The mapping allows us to have an initial 'cut' of Growth Intelligence sectors and products that are relevant for the digital economy, assuming that the SIC classification has some correspondence with the true digital economy, but not enough to capture it accurately.

Step two: We then use a 'threshold rule' to exclude sparse sector and product groups from the analysis. Sparse groups are defined as those with a less an 0.2% share of the Growth Intelligence sector/product within the overall digital economy as conventionally defined. [23]

[22] At this stage of our analysis we only use sector context and product type. Future versions of the research will experiment with using sales type, client type and value chain activity.
[23] We conduct sensitivity checks using 0.1% and 0.5% as variations of the threshold rule. The 0.1% rule would results in a shortlist of 30 sectors and 17 products and it would include 62,167 more observations in the initial four steps (excluding step 3). The 0.5% threshold would result in a shortlist of 16 sectors and 12 products and it would drop extra 37,420 observations. These numbers should be added or subtracted to the number of companies in the digital economy is Table A6, Panel A, as no manual step to identify the digital firms was taken at this stage.

▼ **TABLE A2:** SIC DEFINITION OF DIGITAL ECONOMY COMPANIES BY GROWTH INTELLIGENCE SECTOR

	DESCRIPTION	
	OBSERVATIONS	%
INFORMATION_TECHNOLOGY	101931	54.33
ENTERTAINMENT_FILM_PRODUCTION	17,716	9.44
PUBLISHING	10,847	5.78
TELECOMMUNICATIONS	10,113	5.39
MARKETING_ADVERTISING	9,015	4.81
PRINTING	7,968	4.25
ELECTRICAL_ELECTRONIC_MANUFACTURING	3,889	2.07
RETAIL	3,470	1.85
ARCHITECTURE	2,685	1.43
MECHANICAL_OR_INDUSTRIAL_ ENGINEERING	2,259	1.2
OIL_AND_ENERGY	2,225	1.19
PHOTOGRAPHY	2,040	1.09
NON_PROFIT_ORGANIZATION_MANAGEMENT	1,765	0.94
SEMICONDUCTORS	1471	0.78
APPAREL_AND_FASHION	1,098	0.59
COMPUTER_SOFTWARE	1079	0.58
WHOLESALE	827	0.44
GRAPHIC_DESIGN	677	0.36
INTERNET		494
UTILITIES	436	0.23
WRITING_EDITING	388	0.21
TOTAL	182,393	97

Note: Number of companies that are in the digital economy according to the official classification, by Growth Intelligence sector. The percentage represents the Growth Intelligence sector share within the digital economy used to define the sector relevance.

Source: BIS (2009)

▼ **TABLE A3:** SIC DEFINITION OF DIGITAL ECONOMY COMPANIES BY GROWTH INTELLIGENCE PRODUCT

	DESCRIPTION	
	OBSERVATIONS	%
CONSULTANCY	120,132	64.03
BROADBAND_SERVICES	10,990	5.86
TOYS_GAMES_ACCESSORIES_ORNAMENTS	8,483	4.52
PRINTING_SERVICES	8,032	4.28
PEER_TO_PEER_COMMUNICATIONS	7,364	3.93
CONSUMABLES	7,218	3.85
ELECTRONICS	6,261	3.34
HARDWARE_TOOLS_MACHINERY	5,800	3.09
BOOKS	3,682	1.96
MEDIA_PRODUCTION	2,592	1.38
PHOTOGRAPHY	2,061	1.1
CARE_OR_MAINTENANCE	2,037	1.09
CLOTHING	1,057	0.56
CUSTOM_SOFTWARE_DEVELOPMENT	406	0.22
VEHICLES	402	0.21
TOTAL	186,517	99

Note: Number of companies that are in the digital economy according to the official classification, by Growth Intelligence product. The percentage represents the Growth Intelligence product share within the digital economy used to define the product relevance.

Source: BIS (2009)

For example, as reported in Table A2, more than 50% of the companies in the SIC-defined digital economy are in the Growth Intelligence sector 'information technology'; for this reason, this sector is included in the initial shortlist. This step provides an initial list of 21 'digital' sectors and 16 product groups. Note that these correspond exactly to the existing, 'official' BIS-DCMS list. Table A2 shows the list of sectors and Table A3 shows the list of products we have in the initial cut that uses the threshold of 0.2%. These tables correspond to all the sectors/products that would be included only by using the threshold rule explained above.

Step three: we manually edit the Growth Intelligence sector list to remove irrelevant sectors. We need to do this for two reasons. One is the mis-classification we observe when companies use SIC codes; the other is that SIC codes are a much more detailed classification than GI's codes, therefore some SIC categories are recorded into much broader Growth Intelligence sectors[24].

We adopt our established rules in this step: that is, we exclude sector groups where the only or principal output is non-digital. For instance, we exclude 'oil and energy', 'apparel and fashion', ' non-profit organization management', 'utilities' and 'wholesale'. There are also a number of sectors with companies offering digital and non-digital outputs, such as 'architecture' and 'defence / space'. These are kept in and parsed later, in step six.[25]

[24] For this reason, future research will aim at using further detailed categories (like sales process and client type) for identification, as well as detailed token-level information.

[25] We conduct a sensitivity test on two borderline sector groups: 'retail' and 'mechanical or industrial engineering'. Including those two sectors would results in 373,588 companies in the digital economy, 19.9% of the total economy.

In addition to this, we also include some sectors that would not be included in our initial shortlist, but are where the principle output is likely digital, like 'on-line publishing', 'e-learning' or 'computer network security'.

Step four: the product list is selected in the same way as the sectors, and the rationale is the same. We include product groups like 'digital media', 'photography', 'software mobile application', media delivery'. From the 15 products selected on the base of the threshold we exclude 'care or maintenance', 'toys games accessories ornaments', 'clothing', and 'entertainment experience / activity'[26]. As in the previous step, we again include some products/services that are not in the shortlist, but which describe digital outputs , like 'digital media', 'media delivery' or 'software desktop or server'.

Step five: we make a precision check. As Growth Intelligence sector and product codes are quite broad, we need to look within them using more detailed information. In further versions of the research we will use raw 'token' information to do this. At this stage we use SIC5 codes as a second-best source. Cross-tabulating on SICs reveals two sets of companies that are not clearly digital by our selection rules: 'Activities of head offices' (SIC 70100) and 'Financial intermediation not elsewhere classified' (SIC 64999). We remove companies in these SIC5 categories from our sample.

Step six: in order to be classified as digital, a company has to be in a digital sector and produce a digital product (as defined in the previous steps). Thus, we create a set of companies by 'sector-product', which consists of companies in digital sectors whose principal activity is also digital.

In order to show the importance of steps three-five, we show in Table A4, Panel A, what the economy would look like if we only selected sectors according to steps one, two and six (i.e. without any manual cleaning). As the numbers show, leaving out the manual steps would give us a substantially bigger digital economy – where 25.2% of the companies would be considered digital, compared with around 14% in our preferred estimates.

Panel B shows the resulting economy without the precision step only. We can see that dropping the two SIC sectors does not affect the results massively (i.e. digital firms are 17% of all the firms in the economy), but we strongly think it is important in order to have the neater possible definition.

▼ **TABLE A4:** DIGITAL ECONOMY COUNTS AT DIFFERENT STEPS

	DESCRIPTION	
	OBSERVATIONS	%
A. GROWTH INTELLIGENCE SECTOR AND PRODUCT WITH NO MANUAL ADJUSTMENTS		
OTHER	1,397,647	74.79
DIGITAL ECONOMY	471,120	25.21
B. GROWTH INTELLIGENCE SECTOR AND PRODUCT EXCLUDING STEP 6		
OTHER	1,543,582	82.6
DIGITAL ECONOMY	325,185	17.4
C. GROWTH INTELLIGENCE SECTOR ONLY		
OTHER	1,489,357	79.7
DIGITAL ECONOMY	379,410	20.3
D. GROWTH INTELLIGENCE PRODUCT ONLY		
OTHER	924,116	49.45
DIGITAL ECONOMY	944,651	50.55
TOTAL	1,868,767	100

Note: This table shows the counts of the economy at each step of the procedure used to define the digital economy. Panel A defines the digital economy using Growth Intelligence digital sector by digital product "cells" (like in our preferred definition) when no manual adjustment to the sector included/excluded are made. Panel B excludes only the final precision check. Panel C shows the count if the digital economy was only defined by Growth Intelligence sectors, and Panel D if it was only defined by Growth Intelligence products.

[26] As before, we conduct a sensitivity check on three borderline product groups, 'advertising network', 'consumables' and 'marketplaces'. Including these products will increase the number of digital companies by 11,177, leaving the share of digital economy almost unchanged (up to 14.9%).

A4/ OUTSTANDING ISSUES

We are confident that our cleaning and definition steps give us a clean benchmarking sample of firms, and a workable re-definition of the digital economy in the UK. Of course, every dataset has its limitations, even big data, and there are a number of caveats and outstanding issues.

First, Companies House has unrivalled coverage of the whole of the economy, which for our research questions makes it preferable to datasets such as the BSD. However, even Companies House does not include sole traders, of whom there may be many in parts of the digital economy (such as software developers). Matching these individuals into our data would require matching Companies House data with information from HMRC, which is a substantial exercise and is beyond the scope of this project. Future research could usefully explore these possibilities.

Second, our dataset deals with registered companies, not firms or enterprises (as measured by the BSD). This means that comparisons with most figures provided by Government or the ONS have to be done with some care, as these are typically based on BSD estimates. This is why we have constructed a benchmarking sample that allows SIC/GI comparisons to be made on the same set of companies.

Third, companies are provided with registered addresses, not trading addresses. As we discuss in section 5, this would present problems if companies are typically registered in a small number of locations, or far away from the locations where they typically trade. We minimise the second problem by locating companies in Travel to Work Areas, which roughly correspond to local spatial economies and are likely to capture both registered and trading addresses. The first problem is outstanding, and is visible in some of the detailed local area analysis. In future versions of the research we will be able to deal with it by using trading address information provided by Growth Intelligence.

Fourth, our revenue and employment estimates are necessarily based on sub-samples. While many firms will have no revenue or employment information to report, others have this information and do not provide it to Companies House. We have therefore used diagnostic checks to ascertain whether the set of firms providing revenue and/or employment corresponds to expected characteristics, and is representative of the benchmarking sample as a whole. In both cases our sub-samples pass these checks, but we would still urge some caution in interpreting the results. In future versions of the research we hope to improve employment estimates, in particular.

▼ **TABLE A7:** AVERAGE ANNUAL INFLOW OF FIRMS AND AVERAGE GROWTH, OVER DIFFERENT PERIODS

	A. AVERAGE ANNUAL INFLOW			B. AVERAGE ANNUAL GROWTH RATE		
	1980-1990	1991-2000	2001-2011	1980-1990	1991-2000	2001-2011
SIC 07						
OTHER	9,790	24,375	113,060	7%	9%	13%
DIGITAL ECONOMY	853	3,298	12,888	12%	14%	14%
GI SECTOR AND PRODUCT						
OTHER	9,136	22,730	108,082	7%	8%	13%
DIGITAL ECONOMY	1,507	4,943	17,866	11%	12%	13%

Note: Average annual inflow is the average of the number of firms entering the economy each year. Average annual growth is given by the annual percentage increase of firms over the period.