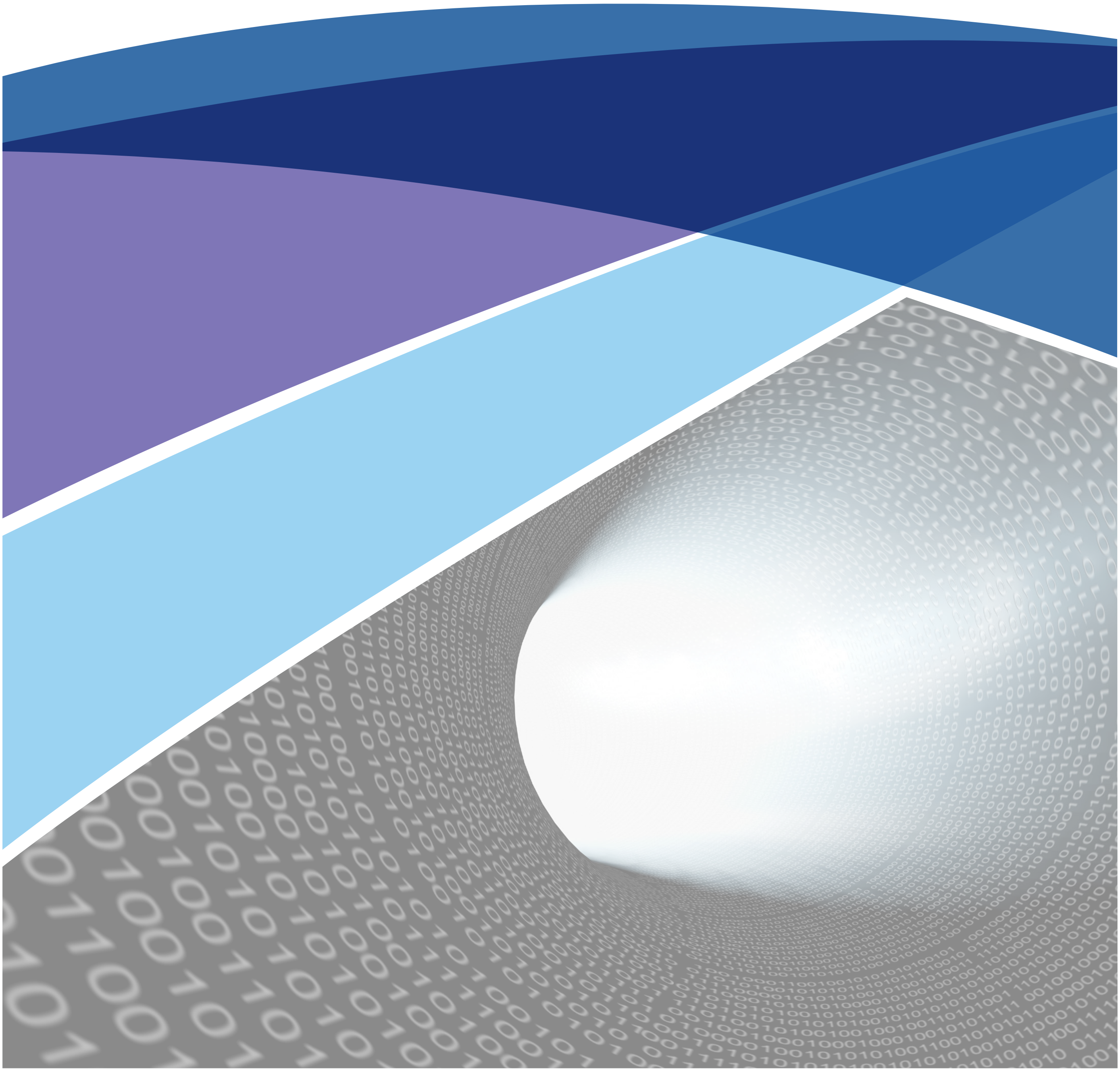
Intellectual
Property
Office

Eight Great Technologies

Big Data

A patent overview



Intellectual Property Office is an operating name of the Patent Office

#8Great

This report was prepared by the
UK Intellectual Property Office Informatics Team
June 2014

e-mail: *informatics@ipo.gov.uk*

Intellectual Property Office
Concept House
Cardiff Road
Newport
NP10 8QQ
United Kingdom

www.ipo.gov.uk/informatics

Contents

1 Introduction

The UK Government has identified ‘eight great technologies’ which will propel the UK to future growth. These are:

- the big data revolution and energy-efficient computing;
- satellites and commercial applications of space;
- robotics and autonomous systems;
- life sciences, genomics and synthetic biology;
- regenerative medicine;
- agri-science;
- advanced materials and nanotechnology;
- energy and its storage.

Patent data can give a valuable insight into innovative activity, to the extent that it has been codified in patent applications, and the IPO Informatics team is producing a series of patent landscape reports looking at each of these technology spaces and the current level of UK patenting on the world stage. As an aid to help people understand the eight great technologies and to consider the direction of future funding, the IPO is offering a comprehensive overview of patenting activity in the each of these technologies.

This report analyses the worldwide patent landscape for technology directed towards big data and its efficient processing. The term “big data” relates to a specific type of data which has such magnitude (typically several petabytes per data set), processing speed requirements and variety that it requires innovative new approaches to its handling and manipulation. Analysis of such data is typically performed via massively parallel computing using, for example, an internetworked collection of computers arranged for cloud-based sharing of the processing. However, the emergence of this specific type of data has been largely fuelled by the recent explosion in social media data, open data and other forms of internet based data, for which a meaningful ten year analysis would not be feasible. Consequently, rather than narrowing this report to just one specific type of data, it has been directed towards patent applications involving the processing any type of large data set(s) for which intensive, distributed processing is required. As such, this report therefore includes consideration of patent applications relating to simulation, modelling and forecasting based upon all types of large data sets, of which “big data” is just one example.

The dataset used for analysis was extracted from worldwide patent databases following detailed discussion and consultation with patent examiners from the Intellectual Property Office who are experts in the field and who, on a day-to-day basis, search, examine and grant patent applications relating to the technologies involved. Throughout the report this data set is referred to as “big data and efficient computing” or simply “big data” for ease of reference[1]. Published patent application data was analysed rather than granted patent data. Applications data gives more information about technological activity than grant data because a number of factors determine whether an application ever proceeds to grant.

[1] For more specific detail of the exact makeup of the dataset see appendix A.6

These include the inherent lag in patent processing at national IP offices worldwide and the patenting strategies of applicants who may file more applications than they ever intend to pursue. In some countries patents are not granted in certain categories such as computer program and mathematical method; this is an issue for areas such as data structures, search algorithms and computer modelling, which are prominent in the big data and efficient computing technology area.

2 Worldwide patent analysis

2.1 Overview

Table 1 gives a summary of the extracted and cleaned dataset used for this analysis of big data and efficient computing technologies. All of the analysis undertaken in this report was performed on this dataset or a subset of this dataset. The worldwide dataset for big data and efficient computing patents published between 2004 and 2013 contains more than 20,000 published patents equating to almost 10,000 patent families. Published patents may be at the application or grant stage, so are not necessarily granted patents. A patent family is one or more published patent originating from a single original (priority) application. Analysis by patent family more accurately reflects the number of inventions present because generally there is one invention per patent family, whereas analysis by raw number of patent publications inevitably involves multiple counting because one patent family may contain dozens of patent publications if the applicant files for the same invention in more than one country. Hence analysis by patent family gives more accurate results regarding the inventive effort that patenting activity represents.

Table 1: Summary of worldwide patent dataset for big data and efficient computing technologies

Number of patent families	9,777		
Number of patent publications	22,421		
Publication year range	2004-2013		
Peak publication year	2013		
Top applicant	IBM Corporation		
Field choices	**Field name**	**Number of entries**	**Coverage**
People	Inventors	17,082	99%
Applicants	Patent assignees	8,231	100%
Countries	Priority countries	32	100%
Technology	IPC sub-group	2,112	99%

Figure 1 shows the total number of published patents by publication year (above) and the total number of patent families by priority year (below – considered to be the best indication of when the original invention took place). Figure 1 suggests a general increase in big data and efficient computing-related patenting over the past decade which continues through 2013. The patent family chart in red does not show any patents filed after 2011 because a patent application is normally published 18 months after the priority date or the filing (application) date, whichever is earlier. Hence, the 2012 and 2013 data is incomplete and has been ignored.

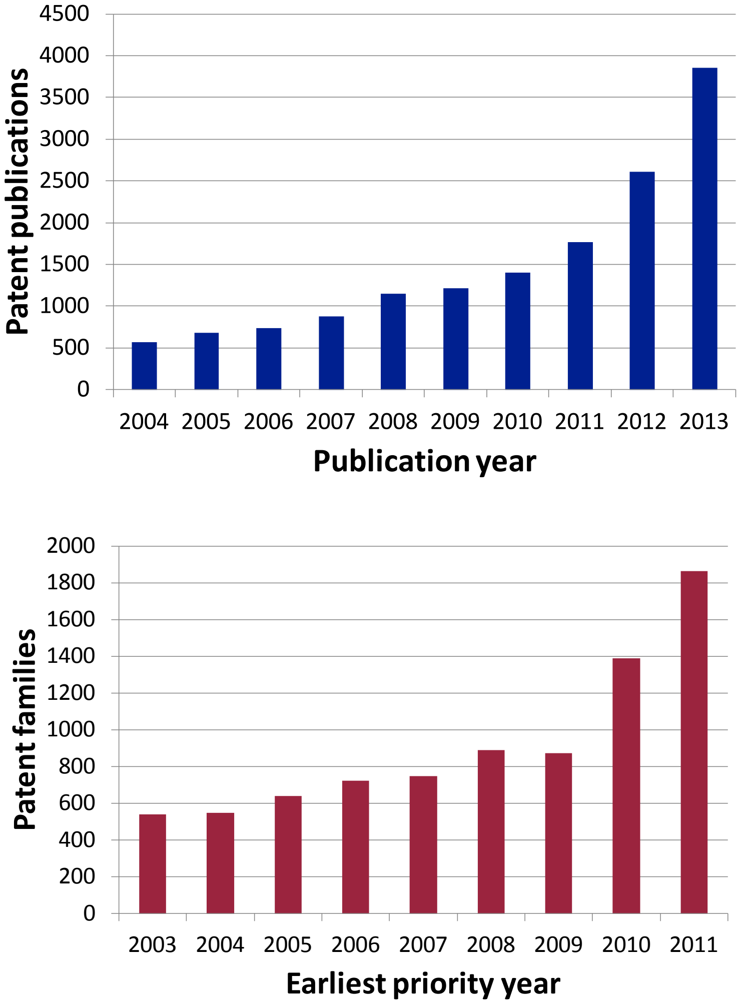


Figure 1: Patent publications by publication year (above) and patent families by priority year (below)

In real-world terms only limited information can be gleaned from the upward trends shown in Figure 1 because general patenting levels globally continue to grow at an ever-increasing rate. Figure 2 addresses this issue by normalising the data shown in Figure 1 and presenting the annual increase in the size of worldwide patent databases across all technologies against the year-on-year increase in the size of the big data and efficient computing dataset. For example, between 2011 and 2012 worldwide patenting across all areas of technology increased by 12.7% and this can be compared to a 47.9% increase in big data and efficient computing patenting over the same time period.

Although Figure 1 shows that there has been a year-on-year increase in patenting in the field of big data and efficient computing over the past decade, Figure 2 shows that this increase has typically been well above the general increase in the size of the worldwide patent databases across all technologies. Across the nine data points shown in Figure 2, patenting activity in big data and efficient computing has been, on average, almost 20% above the year-on-year increase in global patenting activity.

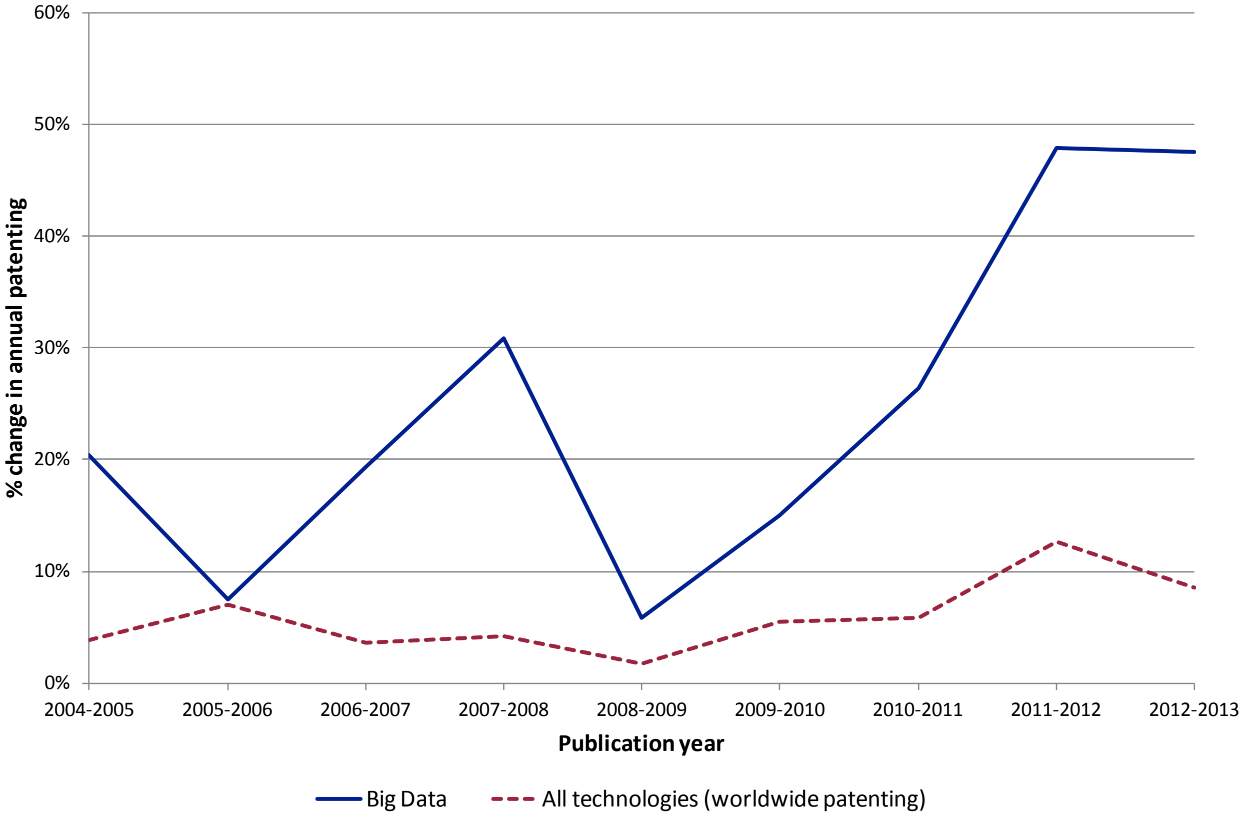


Figure 2: Year-on-year change in big data and efficient computing patenting compared to worldwide patenting across all technologies

Figure 3 shows the priority country distribution across the dataset with more than half of big data and efficient computing patent families having their first filing in the USA. Less than 1% of big data and efficient computing-related patent families are first filed in the UK. Traditionally priority country analysis has been a good indicator of where the invention is actually taking place because many applicants will file patent applications first in the country in which they reside[2], but in recent years drawing firm conclusions from this data is harder because there may be other strategic reasons for an applicant choosing the country of first filing (*e.g.* tax treatment).

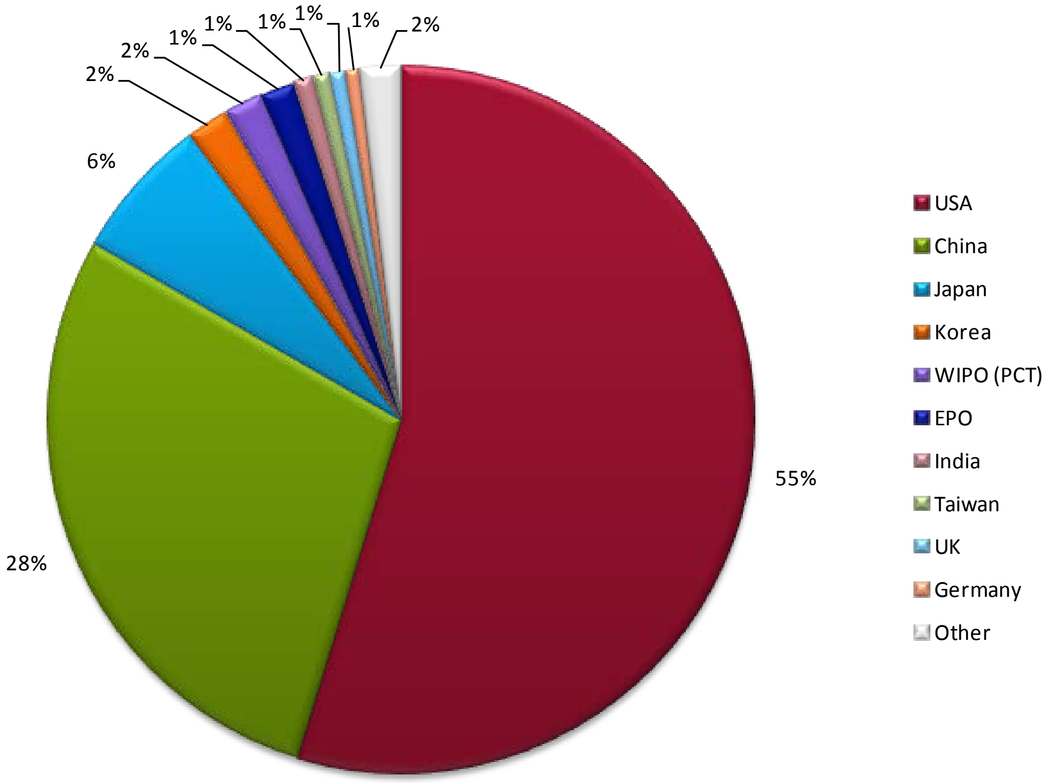


Figure 3: Priority country distribution

When comparing the similarities between the priority country distribution shown in Figure 3 and the inventor country distribution shown in Figure 4 it is important to realise that whilst each family will have only one priority country, it may inventor countries. This may arise for example where collaboration between companies or inventors from different countries has taken place and resulted in a patent that names multiple inventors from different countries.

Figure 3 shows that over 55% of all big data and efficient computing patent families are first filed in the USA, but Figure 4 shows that the USA has less of the overall share when the patent families are distributed by inventor country rather than country of first filing. This may illustrate the strategic importance of the USA, with many inventions made by non-US inventors resulting in priority filings there.

[2] In some countries this is/was a requirement (*e.g.* in the UK this was a requirement until 2005).

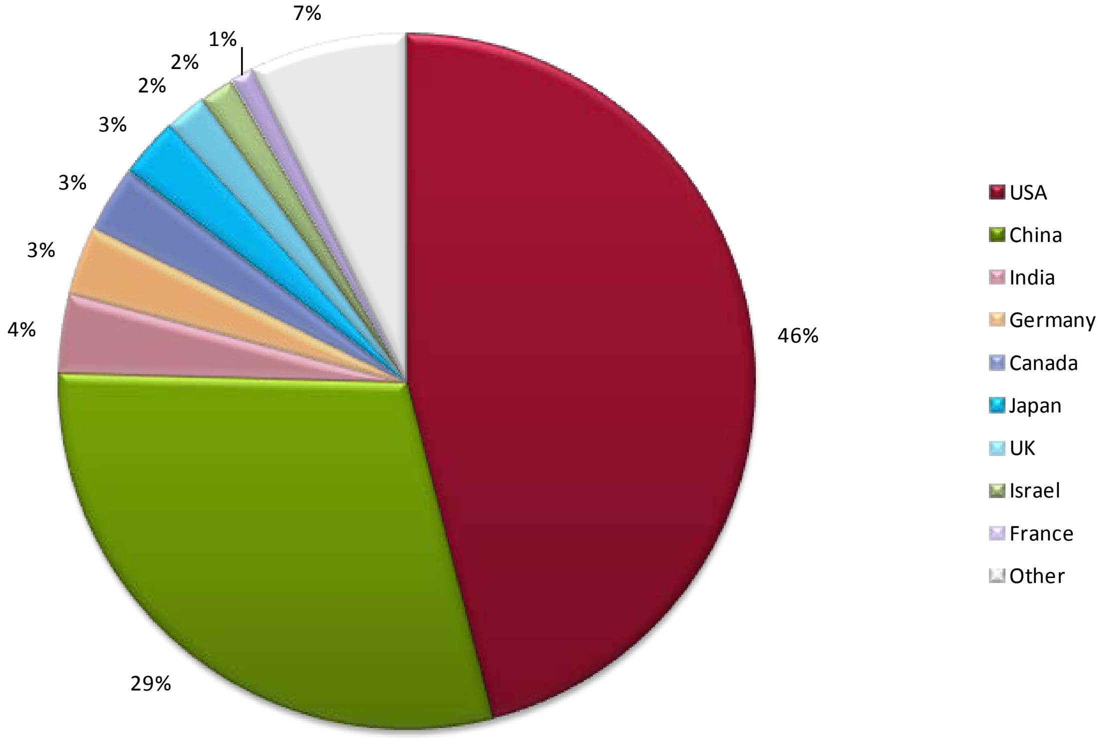


Figure 4: Inventor country distribution

It is well known that there is a greater propensity to patent in certain countries than others, and the trends shown in Figure 4 may change if the figures are corrected for this difference in behaviour. Therefore the Relative Specialisation Index (RSI)[3] for each applicant country (Figure 5) has been calculated to give an indication of the level of invention in big data and efficient computing technologies for each country compared to the overall level of invention in that country.

The RSI shown in Figure 5 appears to suggest a different picture to that shown in Figure 4. The USA and the China are ranked 1st and 2nd in the top inventor countries and appear relatively specialised in the field of big data and efficient computing technologies since their inventors are named on more than two thirds of all big data and efficient computing patent families, but their order is reversed when the RSI is applied as these two countries rank 5th and 4th respectively. They fall below Ireland, Israel, and India. Ireland and Israel do not appear in the top priority countries shown in Figure 3. These high-ranking countries show much greater levels of patenting in this technology space than expected, despite their modest absolute levels of patenting. The UK is ranked 11th with an RSI value of -0.22, suggesting that there are fewer big data and efficient computing patents filed by UK applicants compared to the overall level of patenting from UK applicants across all technology areas. Of course the different conditions for patentability may have an impact here as it should be borne in mind that many of the potential improvements in data processing, particularly with regard to pure business methods and computer software

[3] See Appendix B for full details of how the Relative Specialisation Index is calculated.

routines, are not necessarily protectable by patents[4] and therefore will not be captured by this report.

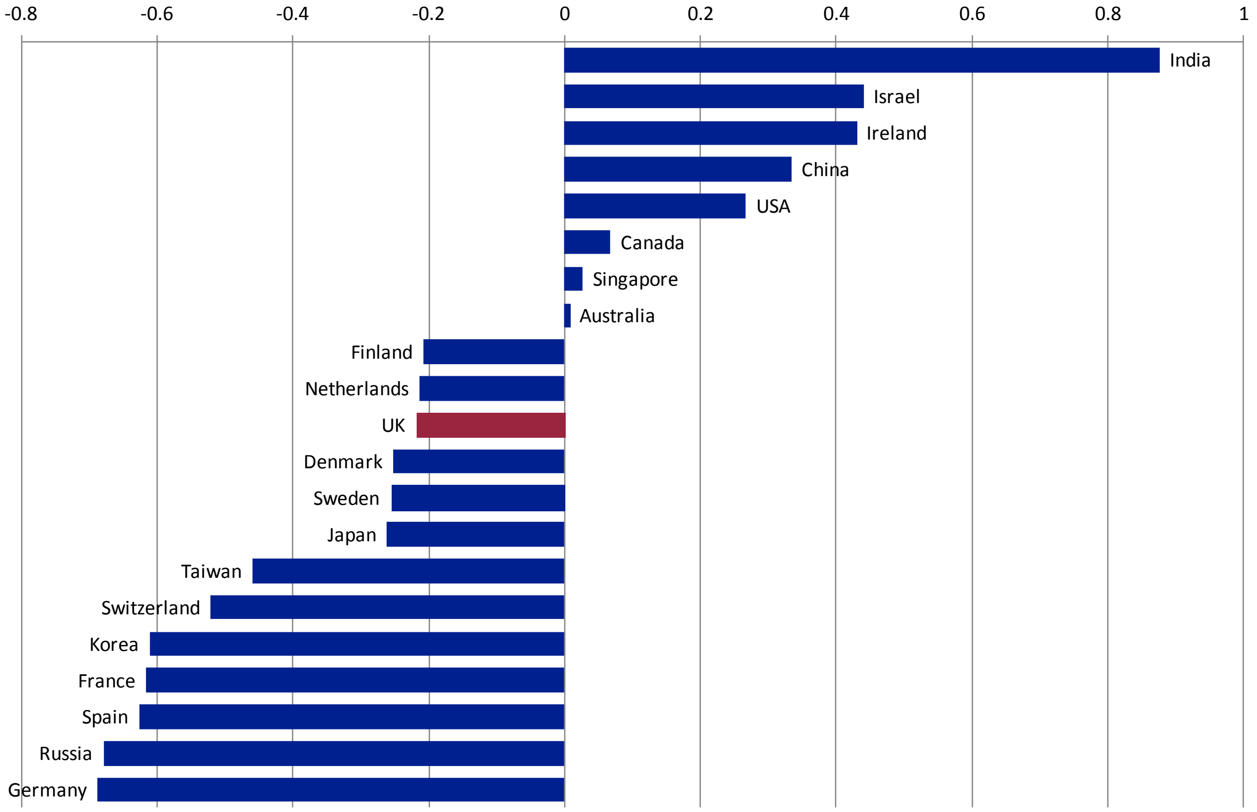


Figure 5: Relative Specialisation Index (RSI) by applicant country

[4] Data processing, particularly with regard to pure business methods and computer software routines, has the potential to fall within of UK's patentability exclusions under s.1(2)(c) of the Patents Act, and similar legal provisions in other patenting authorities. This may have an effect on an applicant's choice file in the UK or Europe rather than in a country which has different law surrounding the patenting of these technologies, for example the USA.

Figure 6 shows the countries in which applicants in the field of big data and efficient computing technologies are interested in seeking patent protection, with the strength of colour reflecting the quantity of published patents in each jurisdiction. The strong showings of Australia, Brazil, Mexico, South Africa and many parts of the European Union, taken in the absence of their appearance in the distributions of priority country and inventor country (Figure 3 and Figure 4), potentially illustrates that though few patents originate from them, these countries are important markets for big data and efficient computing technologies. Published patents filed via the EPO [] and WIPO (PCT) [WIPO] routes are also presented, with Figure 6 showing a relatively strong level of patenting via the EP patent and PCT routes evidenced by the dark orange colour given to the blobs that represent the EPO and WIPO.

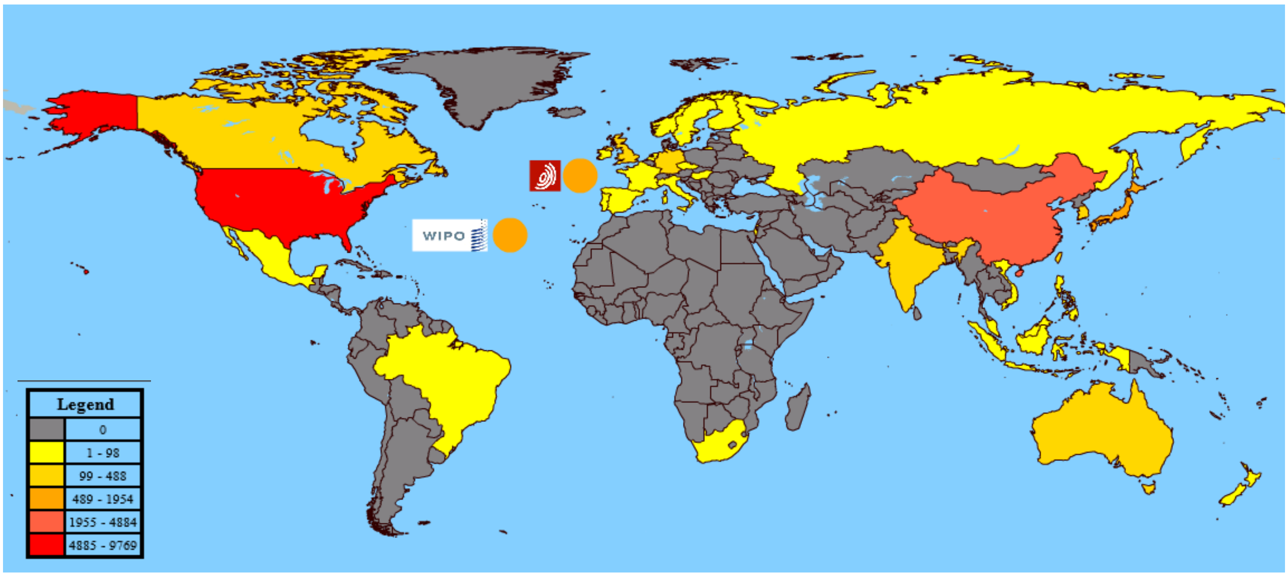


Figure 6: Patent coverage (publication country coverage)

2.2 Top applicants

Patent applicant names within the dataset were cleaned to remove duplicate entries arising from spelling errors, initialisation, international variation and equivalence[5]. Figure 7 shows the top 20 applicants in the dataset. Extensive data cleaning to account for mergers and acquisitions was not undertaken; however SAP and Business Objects have been combined since SAP bought business objects in 2007 and the nomenclature of these patents makes combining them the most sensible data cleaning option.

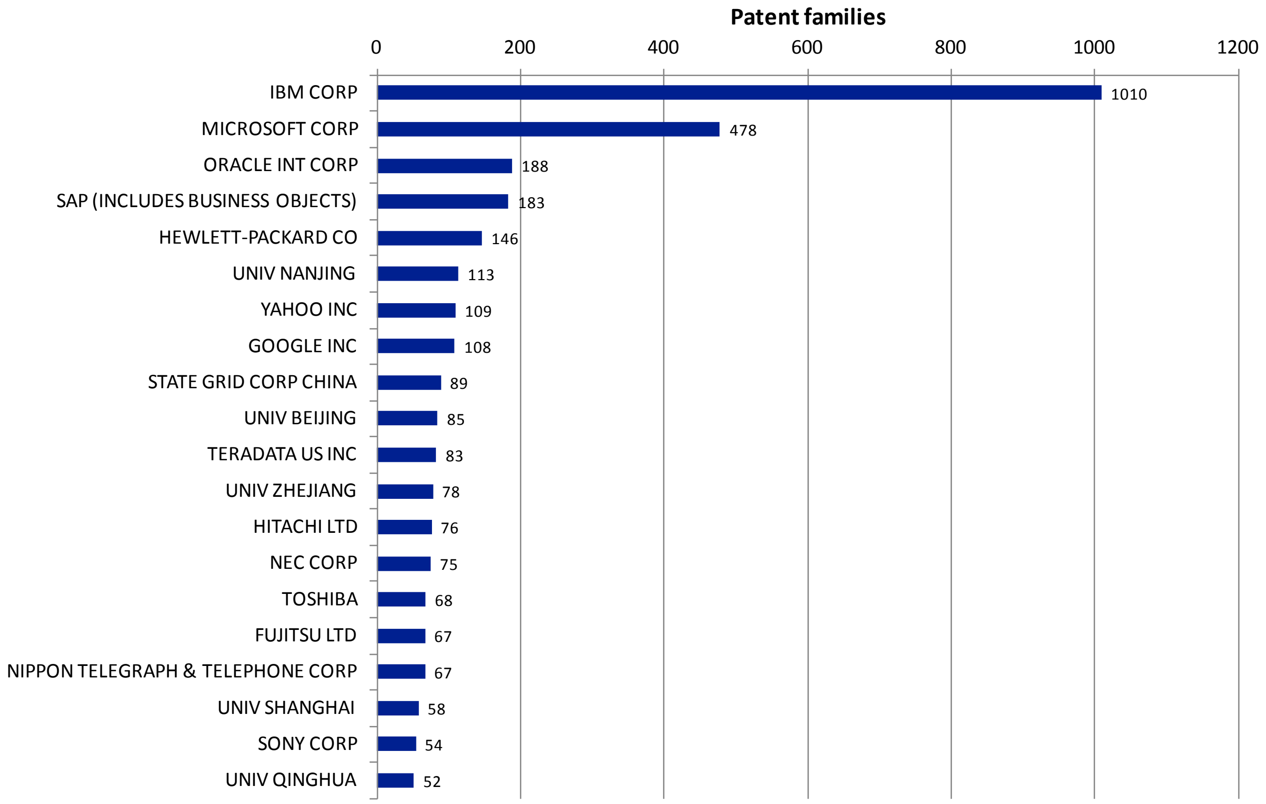


Figure 7: Top applicants

[5] See Appendix A.4 for further details

Figure 8 is a bubble map showing a timeline for the top 20 applicants and shows the filing activity of these applicants in the last 10 years. It shows that most of the top applicants have been involved in big data and efficient computing technologies patenting throughout the last decade in quite a uniform manner. There are, however, some clear exceptions to this uniform trend, most obviously IBM and Microsoft, each of whom have increased their patenting activity significantly in this area over the last decade.

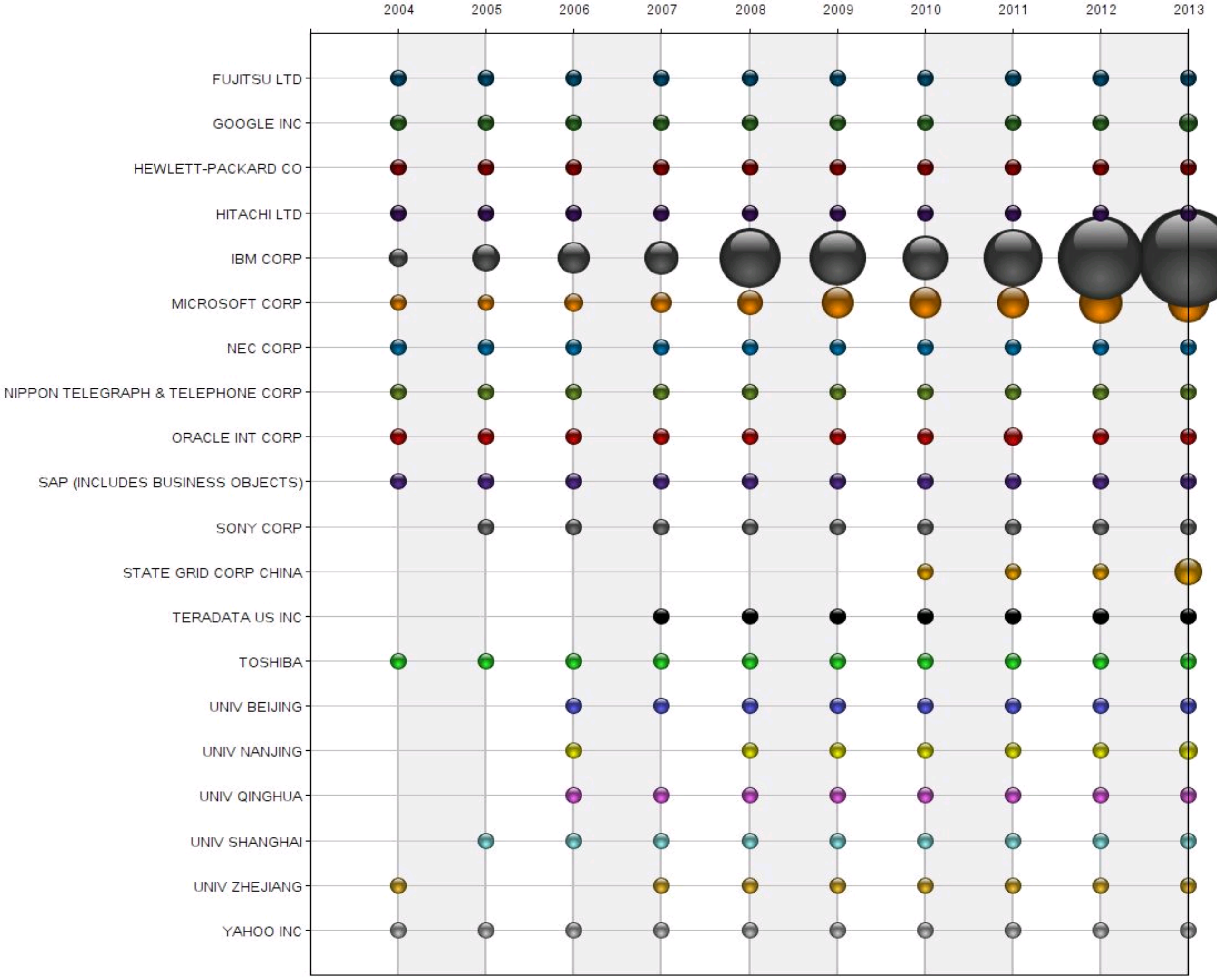


Figure 8: Applicant timeline of published patents by publication year

2.3 Collaboration

Figure 9 is a collaboration map showing all collaborations between the top five applicants in the dataset (the top five shown in Figure 7) and their collaborators. Each dot on the collaboration map represents a patent family and two applicants are linked together if they are named as joint applicants on a patent application. A collaboration map indicates instances where joint work in solving a problem has resulted in a shared application for a patent.

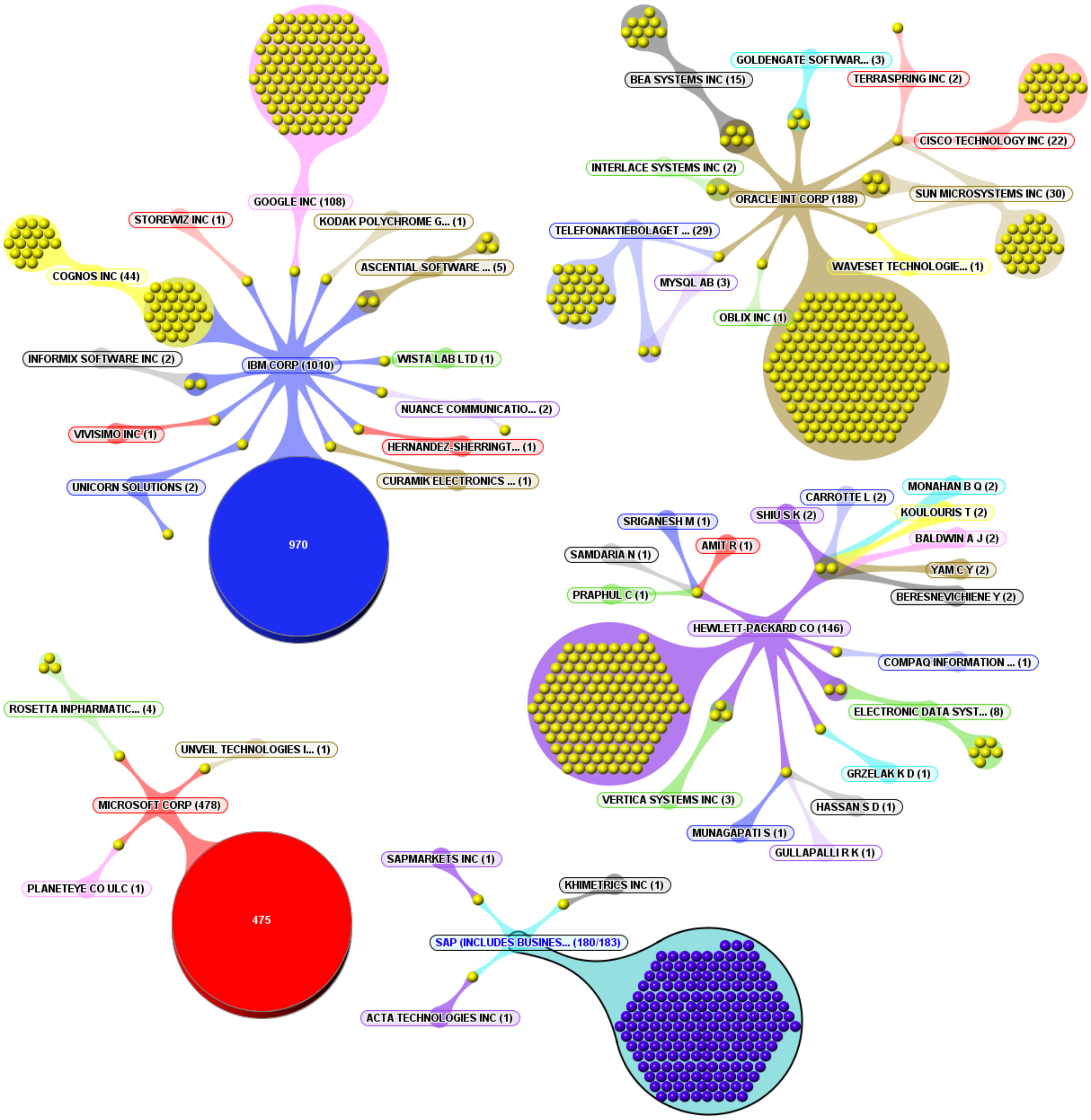


Figure 9: Collaboration map showing all collaborations between the top 5 applicants and their collaborators

Figure 9 shows that none of the top five applicants (IBM, Oracle, Hewlett-Packard, Microsoft and SAP) have worked together on any joint patent applications. Some collaboration is evident, although none of it is with academia and seemingly little is international.

2.4 Technology breakdown

Figure 10 shows the top International Patent Classification (IPC) sub-groups and Table 2 lists the description of each of these sub-groups. The IPC provides for a hierarchical system of language-independent symbols for the classification of patent applications according to the different areas of technology to which they pertain.

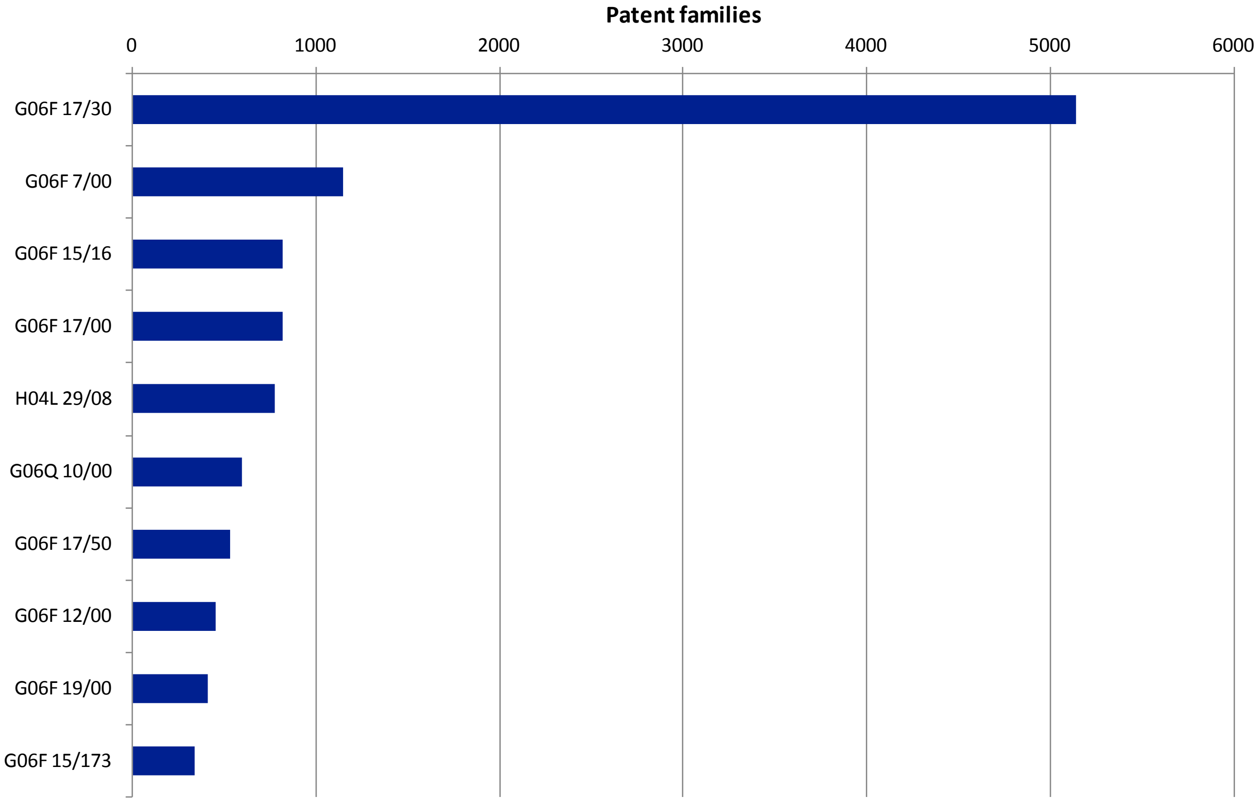


Figure 10: Top IPC sub-groups

Table 2: Key to IPC sub-groups referred to in Figure 10

G06F 17/30	Digital computing or data processing equipment or methods, specially adapted for specific functions -> Information retrieval; Database structures therefor
G06F 7/00	Methods or arrangements for processing data by operating upon the order or content of the data handled
G06F 15/16	Digital computers in general; Data processing equipment in general -> Combinations of two or more digital computers each having at least an arithmetic unit, a programme unit and a register, e.g. for a simultaneous processing of several programmes
G06F 17/00	Digital computing or data processing equipment or methods, specially adapted for specific functions
H04L 29/08	Arrangements, apparatus, circuits or systems, not covered by a single one of groups H04L01/00-H04L27/00 -> Communication control; Communication processing -> characterised by a protocol -> Transmission control procedure, e.g. data link level control procedure
G06Q 10/00	Administration, e.g. office automation or reservations; Management, e.g. resource or project management
G06F 17/50	Digital computing or data processing equipment or methods, specially adapted for specific functions -> Computer-aided design
G06F 12/00	Accessing, addressing or allocating within memory systems or architectures
G06F 19/00	Digital computing or data processing equipment or methods, specially adapted for specific applications
G06F 15/173	Digital computers in general; Data processing equipment in general -> Combinations of two or more digital computers each having at least an arithmetic unit, a programme unit and a register, e.g. for a simultaneous processing of several programmes -> Interprocessor communication -> using an interconnection network, e.g. matrix, shuffle, pyramid, star, snowflake

3 The UK landscape

3.1 Top UK applicants

Figure 11 shows the top UK-based applicants within the big data and efficient computing dataset. The number of patent families shown in the name of IBM, Google and Hewlett-Packard are lower than the values shown in Figure 7 because the data presented in Figure 11 relates to the UK-based parts of these companies. Examples of some of the most recent UK big data and efficient computing patenting from these top UK applicants include: a method for handling cloud computing data according to a ranking value at a virtual machine (IBM); determining query paths by calculated execution costs and interdependence between atoms (British Telecom); a method of indexing multi-dimensional computational fluid dynamics data of aircraft (BAE); a method for enabling cloud service to manage robotic devices (Google); and a method for using microseismic data to characterise natural fracture networks in earth formation (Schlumberger).

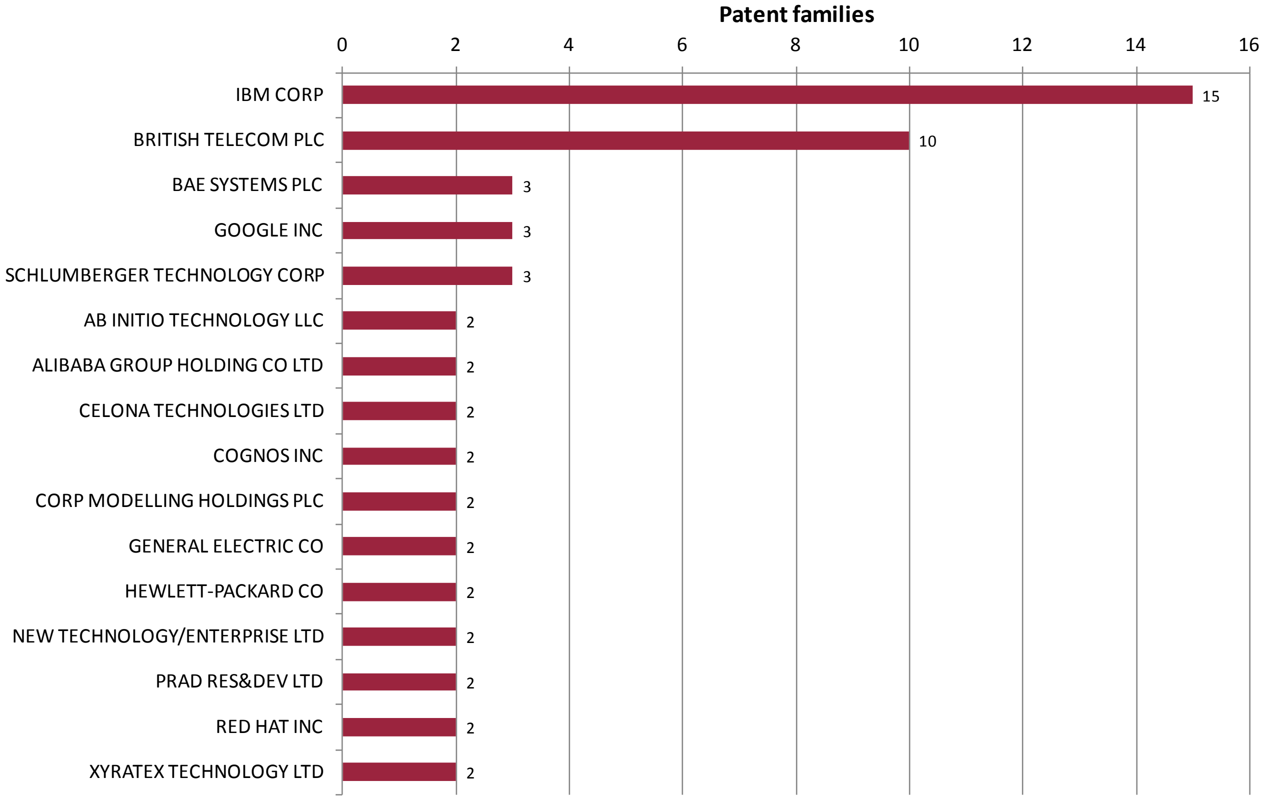


Figure 11: Top UK applicants

3.2 UK inventor mobility

Figure 12 shows the top worldwide applicants with named UK inventors on their published patents. Comparison with the number of patent families from the top UK applicants, Figure 11, confirms that many UK inventors work for UK applicants, including multinational applicants like IBM and Google that have operations in the UK and therefore appear in the top UK applicants chart, Figure 11.

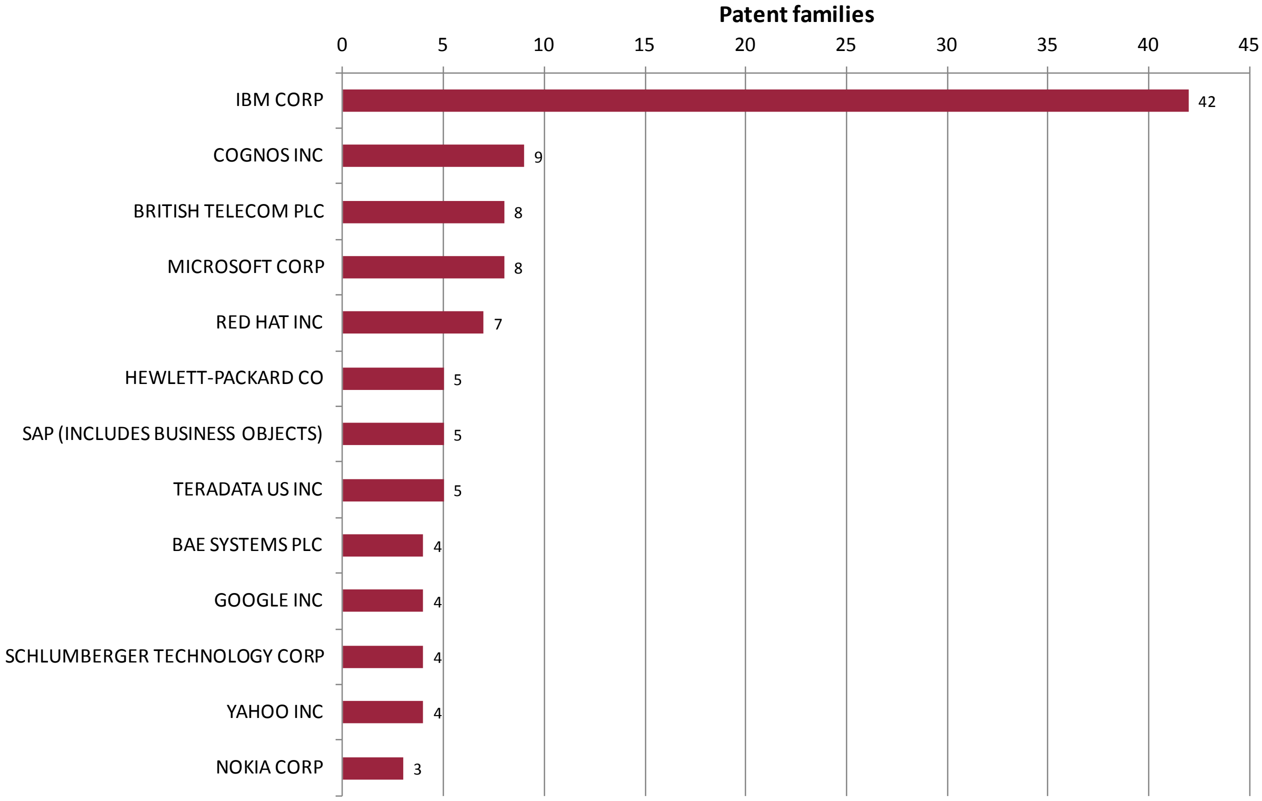


Figure 12: Top worldwide applicants with named UK-based inventors

Of the 97 patent families in the dataset that have UK applicants, only 19 (20%) do not have at least one UK inventor. Conversely, of the 188 patent families that have at least one UK inventor, 110 (59%) do not have a UK applicant. This may suggest that UK inventors are highly sought after in this technology area: even though UK applicants usually employ UK inventors, 59% of UK inventors are employed by non-UK patent applicants.

3.3 How active is the UK?

A subset of the main worldwide patent dataset designed to reflect UK patenting activity was selected[6]. Figure 13 shows the annual change in big data and efficient computing patenting arising from UK patenting activity against the worldwide year-on-year change in this field shown in Figure 2; this shows that UK patenting activity grew considerably most years between 2004 and 2013.

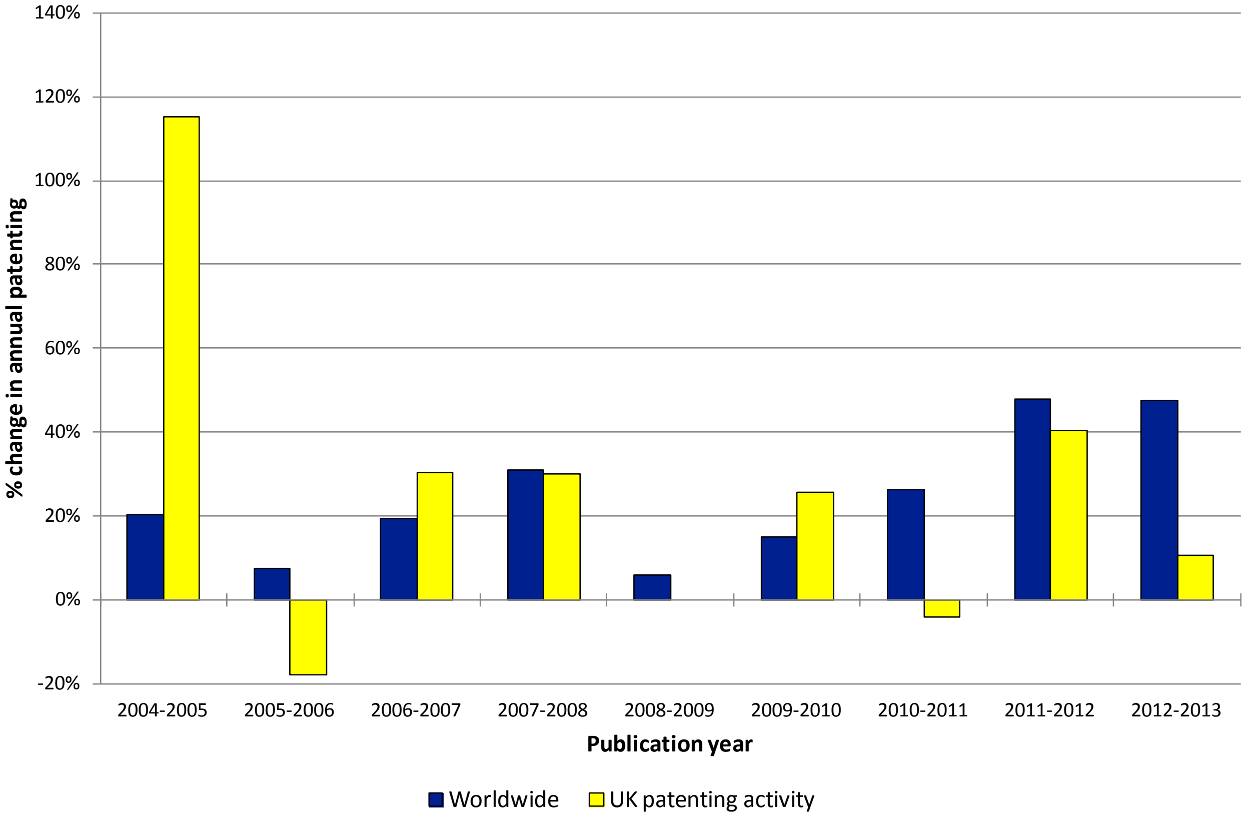


Figure 13: Year-on-year change in UK and worldwide patenting

[6] This was achieved by taking all patent families which have a GB inventor, GB applicant or a GB priority country.

Similar patent subsets were created to reflect patenting activity taking place in several comparator countries (France, Germany, USA, Japan and China) to produce the comparison chart shown in Figure 14.

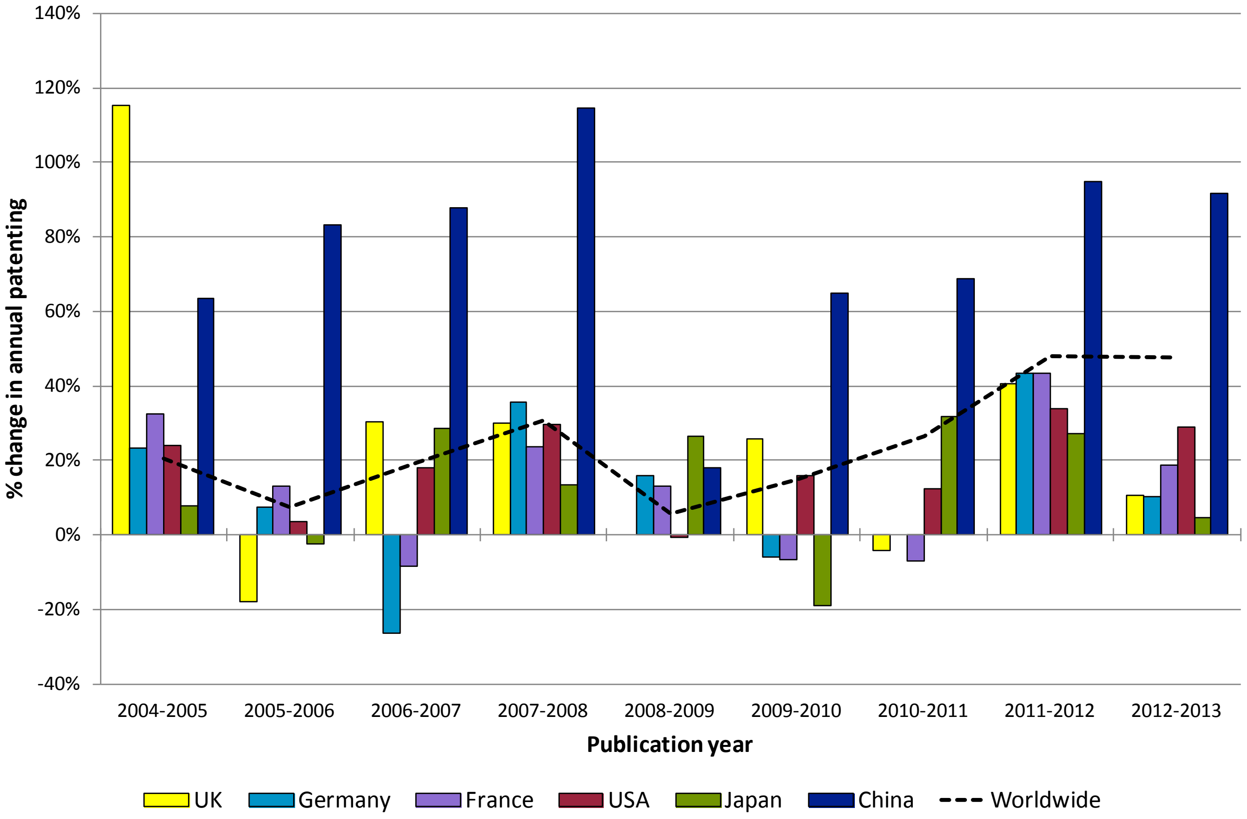


Figure 14: Year-on-year change in UK big data and efficient computing patenting against comparison countries

Chinese patenting activity dominates across most of the time period analysed, with a more than 60% increase in patenting activity in every year other than 2008-2009. In 2004 Chinese patenting activity resulted in just 11 patent families compared to 1632 in 2013 and the average annual growth of Chinese patenting activity in big data and efficient computing technologies over the time period measured is more than 75%. This significant and rapid Chinese patenting activity is not specific to big data and efficient computing technologies and is often seen in a wide range of different technology spaces.

Excluding the first period 2004-2005, for which the large increase can be attributed to an initial low-level of patenting, UK patenting activity in big data and efficient computing has, on the whole, increased over recent years and the year-on-year changes are comparable to the growth seen in Germany, France and Japan.

4 Patent landscape map analysis

In order to give a snapshot as to what the patent landscape looks like for this technology space, a patent map provides a visual representation of the dataset. Published patents (not patent families) are represented on a patent map by dots and the more intense the concentration of patents (*i.e.* the more closely related they are) the higher the topography as shown by contour lines. The patents are grouped according to the occurrence of keywords in the title and abstract and examples of the reoccurring keywords appear on the patent map[7].

Figure 15 shows a patent landscape map of the most recent five year period for big data and efficient computing technologies (publication years 2009-2013). The largest 'snow-capped peaks' around the centre of the map show that the highest concentration of patents in this dataset relate to patents comprising keywords such as "clusters"/"process"/"efficiently", "image"/"processing"/"effective", "personal"/"device"/"digital assistant" and "useful"/"identifying"/"provides", which suggest that these most prolific areas of patenting are directed towards processing of data using clusters of processors, image processing, trend identification and providing personal digital assistant devices with the ability to handle large data sets (typically by offloading memory, power and processor intensive tasks to more powerful, remote devices, unconstrained by the battery limitations of a smartphone or tablet).

[7] Further details regarding how patent landscape maps are produced is given in Appendix C.

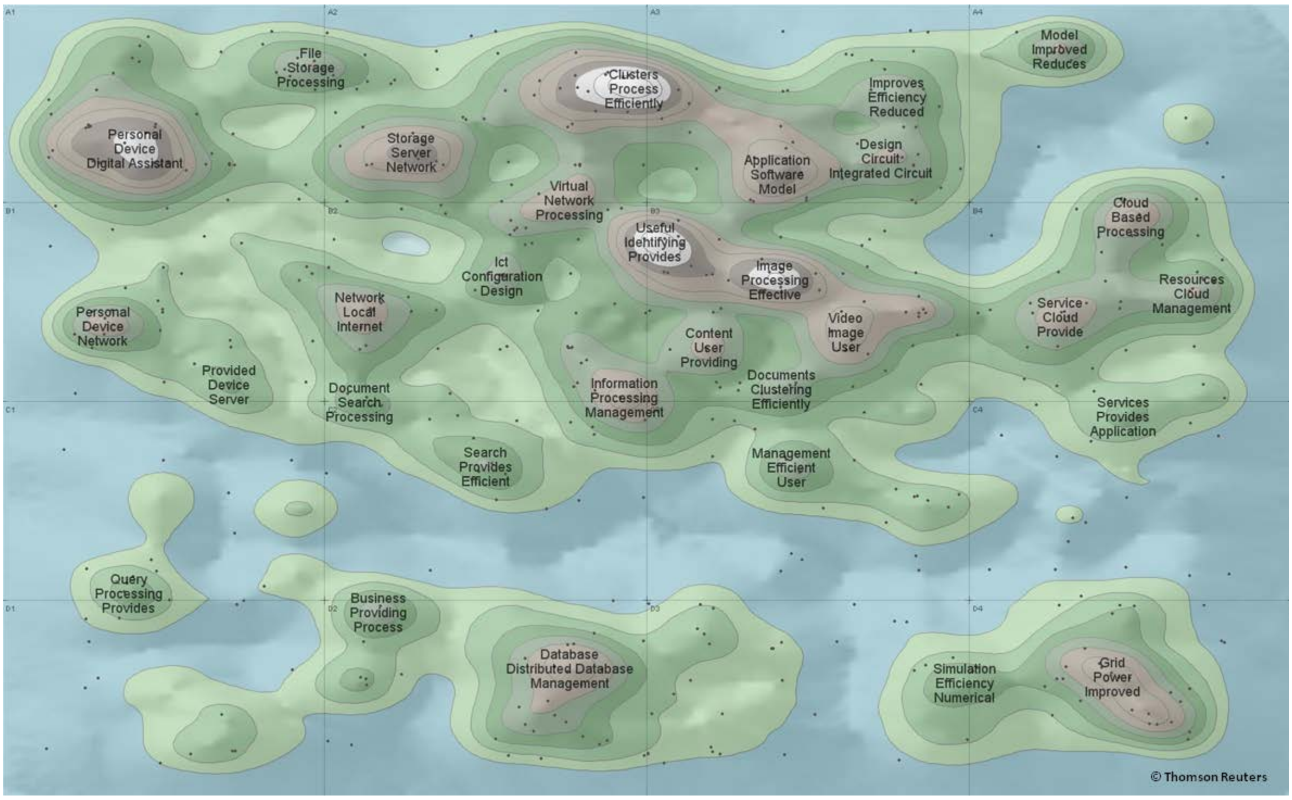

A1
A2
A3
A4
B1
B2
B3
B4
C1
C4
D1
D2
D3
D4
Model Improved Reduces
File Storage Processing
Clusters Process Efficiently
Improves Efficiency Reduced
Personal Device Digital Assistant
Storage Server Network
Application Software Model
Design Circuit Integrated Circuit
Virtual Network Processing
Cloud Based Processing
Useful Identifying Provides
Ict Configuration Design
Image Processing Effective
Resources Cloud Management
Network Local Internet
Personal Device Network
Service Cloud Provide
Video Image User
Content User Providing
Provided Device Server
Documents Clustering Efficiently
Document Search Processing
Information Processing Management
Services Provides Application
Search Provides Efficient
Management Efficient User
Query Processing Provides
Business Providing Process
Database Distributed Database Management
Simulation Efficiency Numerical
Grid Power Improved
© Thomson Reuters

Figure 15: Patent landscape map of all patents relating to big data (2009-2013)

The patent landscape map shown in Figure 16 is the same patent map shown in Figure 15, but with specific patents (dots) highlighted. The map in Figure 16 highlights the patents filed by the top five worldwide applicants (as shown in Figure 7) between 2009 and 2013. Since these patent landscape maps are produced using all patent publications rather than patent families, very tight clusters of several patents are likely to be from the same applicant and relate to one patent family (invention) rather than several similar, but separate inventions.

Figure 16 shows that most of the top worldwide applicants have a fairly broad spread of interests across the technology space with a range of big data patents across the majority of the patent landscape map. Worthy of note however is Microsoft, who appears to have a particular focus on personal device networks and personal digital assistants (see “person”/”device”/”network” and “personal”/”device”/”digital assistant” peaks); Oracle who appears to operate mostly in the fields of distributed databases and storage servers (“database”/”distributed database”/”management” peak and activity around “storage”/”server”/”network” peaks); and SAP with their business data processing technologies (“business”/”providing”/”process” peak).

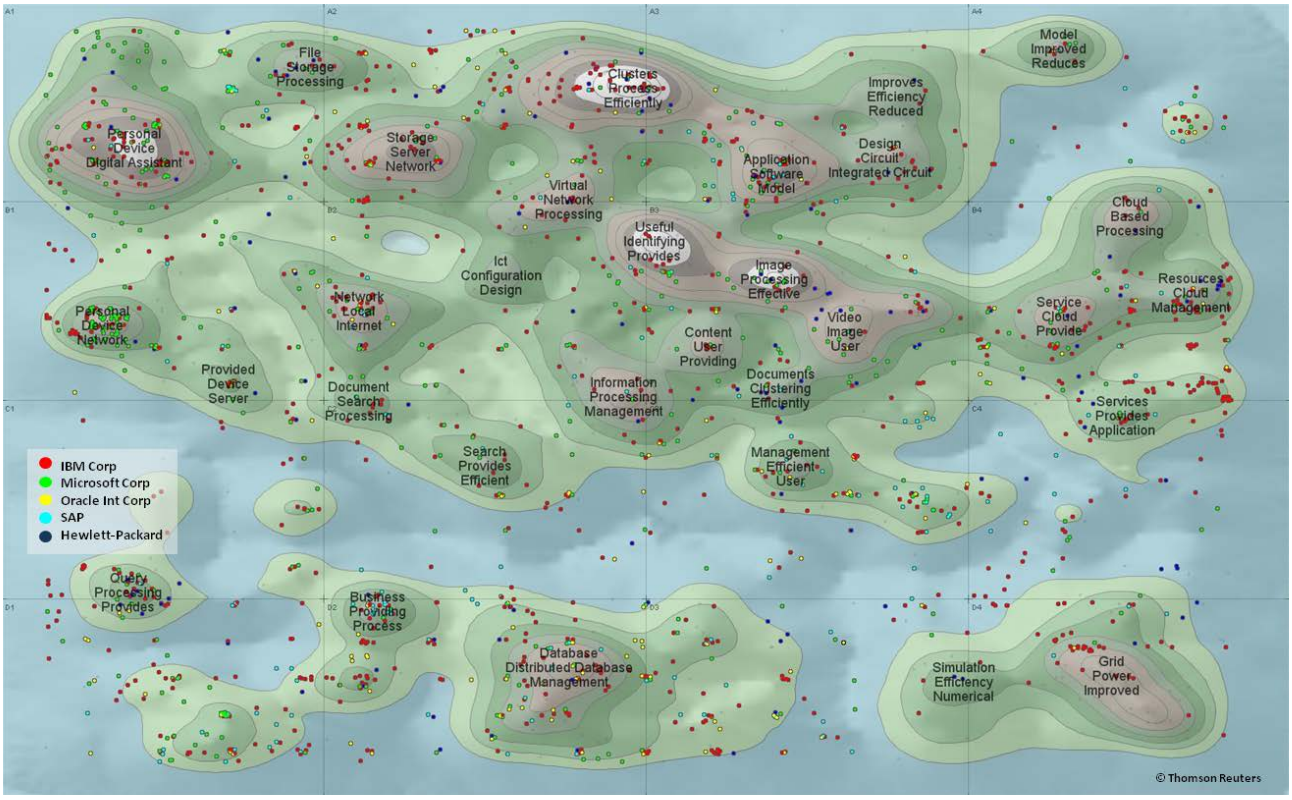

File Storage Processing
Clusters Process Efficiently
Model Improved Reduces
Improves Efficiency Reduced
Personal Device Digital Assistant
Storage Server Network
Application Software Model
Design Circuit Integrated Circuit
Virtual Network Processing
Cloud Based Processing
Useful Identifying Provides
Ict Configuration Design
Image Processing Effective
Resources Cloud Management
Personal Device Network
Network Local Internet
Service Cloud Provide
Video Image User
Content User Providing
Provided Device Server
Document Search Processing
Information Processing Management
Documents Clustering Efficiently
Services Provides Application
Search Provides Efficient
Management Efficient User
● IBM Corp
● Microsoft Corp
● Oracle Int Corp
● SAP
● Hewlett-Packard
Query Processing Provides
Business Providing Process
Database Distributed Database Management
Simulation Efficiency Numerical
Grid Power Improved
© Thomson Reuters

Figure 16: Patent landscape map with top 5 worldwide applicants highlighted

The same patent landscape map has been has been split by publication year, Figure 17, with patents published in 2009 shown in green and patents published in 2013 shown in red. This shows the areas of patenting activity in the first and last years of the analysable publication date range (2009-2013) and highlights the patenting shift into new areas. The two most noticeable areas of increased patenting activity in 2013 compared to 2009 relate to data processing improvements for personal digital assistants and to the field of cloud based processing and have been highlighted in yellow. When taken in conjunction with Figure 16, it is clear that Microsoft and IBM are most heavily involved in these growth areas. Also of note is the peak labelled “grid”/”power”/”improved” (circled orange) which would appear to relate to simulation and analysis of power grids. This area of technology on the landscape map is heavily dominated by State Grid Corp China though, as can be seen from Figure 16 IBM is also active in this field.

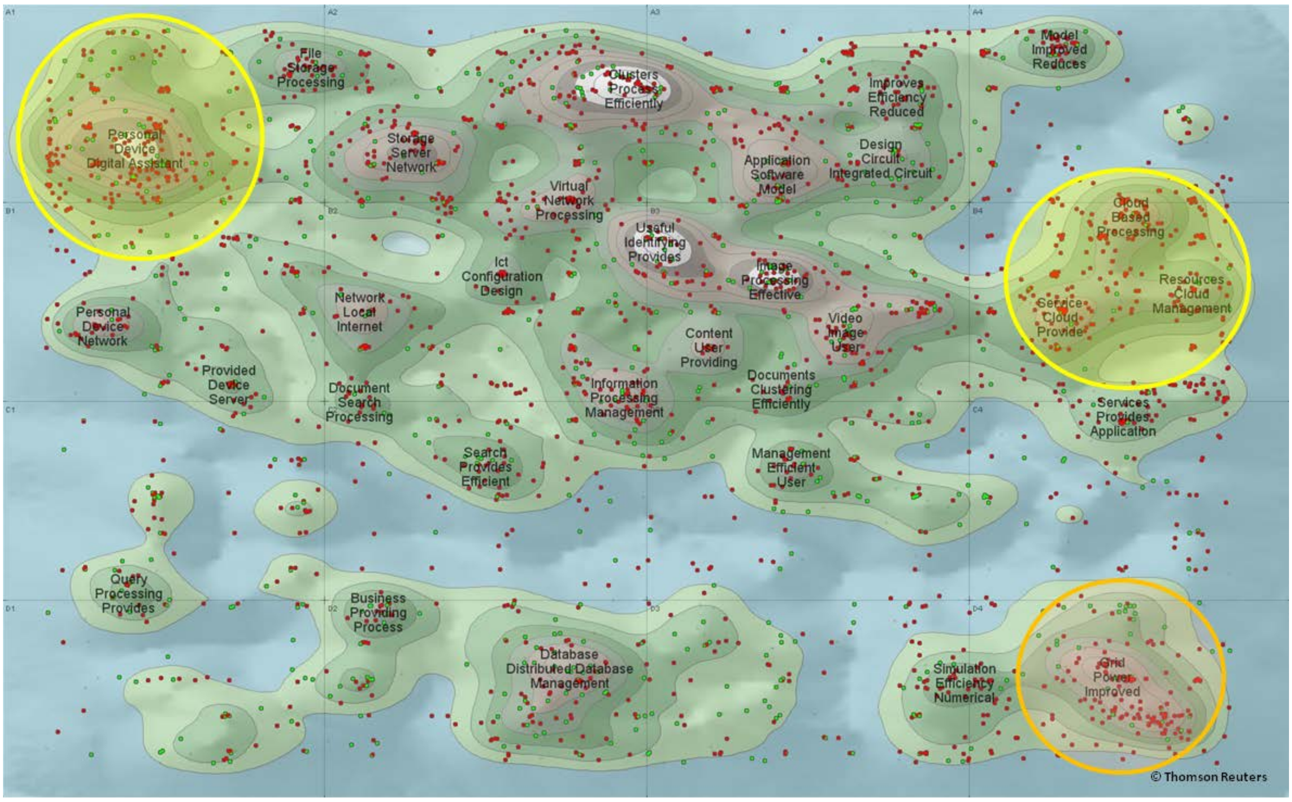

File Storage Processing
Clusters Process Efficiently
Model Improved Reduces
Improves Efficiency Reduced
Personal Device Digital Assistant
Storage Server Network
Application Software Model
Design Circuit Integrated Circuit
Virtual Network Processing
Useful Identifying Provides
Cloud Based Processing
Ict Configuration Design
Image Processing Effective
Resources Cloud Management
Network Local Internet
Personal Device Network
Content User Providing
Video Image User
Service Cloud Provide
Provided Device Server
Document Search Processing
Information Processing Management
Documents Clustering Efficiently
Services Provides Application
Search Provides Efficient
Management Efficient User
Query Processing Provides
Business Providing Process
Database Distributed Database Management
Simulation Efficiency Numerical
Grid Power Improved
© Thomson Reuters

Figure 17: Patent landscape map landscape map split by publication year (2009 in green; 2013 in red)

5 Conclusions

There are more than 22,000 published patent applications between 2004 and 2013 relating to big data and efficient computing technologies, resulting in almost 10,000 patent families. Patenting activity in this field has grown steadily over the last decade and has seen its highest increases in annual patenting over the last two years (2011-2012 and 2012-2013) of the present data set. The growth has continually been above the general worldwide increase in patenting, showing a small increase of 0.4% over worldwide patenting for the 2005-2006 period and showing a maximum increase of 39% for 2012-13.

IBM has the most patent families (inventions), with more than double those of its nearest competitor, Microsoft. SAP AG, with its acquisition of Business Objects (USA) and their patent portfolio, represents the highest placed European applicant, whilst IBM also heads up the list of top UK applicants. IBM's prominence in the UK patent filing field should come as no surprise since has around 20,000 UK employees including around 3,000 at the IBM research and development lab in Winchester. British Telecom PLC has the second largest number of UK patent families.

80% of all big data and efficient computing patent families (inventions) are filed by US and Chinese applicants, with UK applicants accounting for just 1.2% of the dataset and filing slightly fewer big data and efficient computing patents than expected given the overall level of patenting activity from UK applicants across all areas of technology. Against this, however, it should be borne in mind that many of the potential improvements in data processing, particularly with regard to pure business methods and computer software routines, are not necessarily protectable by patents and therefore will not be captured by this report.

UK patenting activity in big data and efficient computing has, on the whole, increased over recent years and the year-on-year changes are comparable to the growth seen in Germany, France and Japan.

Appendix A Interpretation notes

A.1 Patent databases used

The *Thomson Reuters* World Patent Index (WPI) was interrogated using *Thomson Innovation*[8], a web-based patent analytics tool produced by *Thomson Reuters*. This database holds bibliographic and abstract data of published patents and patent applications derived from the majority of leading industrialised countries and patent organisations, *e.g.* the World Intellectual Property Organisation (WIPO), European Patent Office (EPO) and the African Regional Industry Property Organisation (ARIPO). It should be noted that patents are generally classified and published 18 months after the priority date. This should be borne in mind when considering recent patent trends (within the last 18 months).

The WPI database contains one record for each patent family. A patent family is defined as all documents directly or indirectly linked via a priority document. This provides an indication of the number of inventions an applicant may hold, as opposed to how many individual patent applications they might have filed in different countries for the same invention.

A.2 Priority date and publication date

Priority date: The earliest date of an associated patent application containing information about the invention.

Publication date: The date when the patent application is published (normally 18 months after the priority date or the application date, whichever is earlier).

Analysis by priority year gives the earliest indication of invention.

A.3 WO and EP patent applications

International patent applications (WO) and European patent applications (EP) may be made through the World Intellectual Property Organization (WIPO) and the European Patent Office (EPO) respectively.

International patent applications may designate any signatory states or regions to the Patent Cooperation Treaty (PCT) and will have the same effect as national or regional patent applications in each designated state or region, leading to a granted patent in each state or region.

European patent applications are regional patent applications which may designate any signatory state to the European Patent Convention (EPC), and lead to granted patents having the same effect as a bundle of national patents for the designated states.

[8] http://info.thomsoninnovation.com

Figures for patent families with WO and EP as priority country have been included for completeness although no single attributable country is immediately apparent.

A.4 Patent documents analysed

The satellite patent dataset for analysis was identified in conjunction with patent examiner technology-specific expertise. A search strategy was developed and the resulting dataset was extracted in June 2014 using International Patent Classification (IPC) codes, Co-operative Patent Classification (CPC) codes and keyword searching of titles and abstracts in the *Thomson Reuters* World Patent Index (WPI) and limited to patent families with publications between 2004 and 2013.

The applicant and inventor data was cleaned to remove duplicate entries arising from spelling errors, initialisation, international variation (Ltd, Pty, GmbH *etc*.), or equivalence (Ltd., Limited, *etc*.).

A.5 Analytics software used

The main computer software used for this report is a text mining and analytics package called *VantagePoint*[9] produced by *Search Technology* in the USA. The patent records exported from *Thomson Innovation* were imported into *VantagePoint* where the data is cleaned and analysed. The patent landscape maps used in this report were produced using *Thomson Innovation*.

A.6 Search strategy

The dataset used for this report was obtained using following keywords, which were used in conjunction with each other and with the IPC and CPC terms below:

Keywords:

“big data”, Hadoop®, Yarn, Aster®, Datameer®, FICO® Blaze, Vertica®, Platfora®, Splunk®, MapReduce, “open data”, “data warehous*”, informatic*, “data mine?”, “data mining”, simulate*, model*, analy*, “artificial intelligence”, “neural network*”, “distributed *, (cluster*, cloud*, grid?) [within 3 words of] (based, comput*, server?, process*, software, application), croudsourc*, “crowd sourc*”, “massively parallel process*”, “massively parallel software”, “massively parallel database?”, “distributed process*”, “distributed server?”, “distributed quer*”, “distributed database?”, “massive data”

CPC/IPC:

G06F(3/0625, 9/5072, 17/30*, 17/30147, 17/30283, 17/50*, 17/30539, 17/30545, 17/30557, 17/3056*, 17/30572, 17/30575, 17,30592, 17/30598, 17/30601, 19/1*, 19/30*, 19/70*)

G06Q (10/6*, 30/0201, 30/0202)

G06N (3*, 5*, 7*, 99/005)

[9] http://www.thevantagepoint.com

Appendix B Relative Specialisation Index

Relative Specialisation Index (RSI) was calculated as a correction to absolute numbers of patent families in order to account for the fact that some countries file more patent applications than others in all fields of technology. In particular, US and Japanese inventors are prolific patentees. RSI compares the fraction of satellite patents found in each country to the fraction of patents found in that country overall. A logarithm is applied to scale the fractions more suitably. The formula is given below:

$$\log_{10}\left(\frac{{}^{n_i}/_{n_{total}}}{{}^{N_i}/_{N_{total}}}\right)$$

where

n_i = number of big data patent publications in country i

n_{total} = total number of big data patent publications in dataset

N_i = total number of patent publications in country i

N_{total} = total number of patent publications in dataset

The effect of this is to highlight countries which have a greater level of patenting in satellites than expected from their overall level of patenting, and which would otherwise languish much further down in the lists, unnoticed.

Appendix C Patent landscape maps

A patent landscape map is a visual representation of a dataset and is generated by applying a complex algorithm with four stages:

i) **Harvesting documents** – When the software harvests the documents it reads the text from each document (ranging from titles through to the full text). Non-relevant words, known as stopwords, (*e.g.* “a”, “an”, “able”, “about” *etc*) are then discounted and words with common stems are then associated together (*e.g.* “measure”, “measures”, “measuring”, “measurement” *etc*). For the purpose of this big data and efficient computing report, the harvesting involved reading text from the “DWPI Advantages” and “DWPI Uses” fields of the Derwent WPI database since this enabled mapping of the analysis of documents with foreign language abstracts without removing those documents from the landscape and provided a more meaningful grouping of the technologies.

ii) **Analysing documents** – Words are then analysed to see how many times they appear in each document in comparison with the words’ frequency in the overall dataset. During analysis, very frequently and very infrequently used words (*i.e.* words above and below a threshold) are eliminated from consideration. A topic list of statistically significant words is then created.

iii) **Clustering documents** – A Naive Bayes classifier is used to assign document vectors and Vector Space Modelling is applied to plot documents in n-dimensional space (*i.e.* documents with similar topics are clustered around a central coordinate). The application of different vectors (*i.e.* topics) enables the relative positions of documents in n-dimensional space to be varied.

iv) **Creating the patent map** – The final n-dimensional model is then rendered into a two-dimensional map using a self-organising mapping algorithm. Contours are created to simulate a depth dimension. The final map can sometimes be misleading because it is important to interpret the map as if it were formed on a three-dimensional sphere.

Thus, in summary, published patents are represented on the patent map by dots and the more intense the concentration of patents (*i.e.* the more closely related they are) the higher the topography as shown by contour lines. The patents are grouped according to the occurrence of keywords in the title and abstract and examples of the reoccurring keywords appear on the patent map. Please remember there is no relationship between the patent landscape maps and any geographical map.

Please note that the patent maps shown in this report are snapshots of the patent landscape, and that patent maps are best used an interactive tool where analysis of specific areas, patents, applicants, inventors *etc* can be undertaken ‘on-the-fly’.

Intellectual Property Office

Concept House
Cardiff Road
Newport
NP10 8QQ
United Kingdom

www.ipo.gov.uk

#8Great