New Data for Innovation Policy

Hasan Bakhshi and Juan Mateos-Garcia, Nesta

DRAFT

August 2016

Abstract

We contend that existing datasets based on survey or administrative data queried via standard industrial and occupational standards are ill suited for the purposes of innovation policy, and therefore, of its ability to attain its goals - to address market, system and emergence failures preventing new ideas from being applied. These datasets are constrained in their ability to identify novel sectors, map innovation networks, characterise complex ecosystems, and operationally target innovation interventions. New data – big data, web data and open data – data combinations and interactive mapping and reporting tools could help address some of the limitations. We review recent studies that have used these methods, and our own experience with them. We also discuss challenges – related to data quality and trust, relevance and the need to develop complementary capabilities and processes in government, which stand on the way of their more extensive application and impact. We hypothesise that although new data can greatly improve the effectiveness of innovation policy, some defining features of this policy domain – low volume of decisions, and high levels of uncertainty – may limit its transformational impact compared to other policy areas.

1) Introduction

Policymakers in various parts of government, from financial regulation to policing and education to transport, are exploring new forms of data and new ways of analysing it, anticipating that it will lead to better public policies. The hope is that data analytics – the analysis of big data ('high volume, high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization' (Diebold, 2012)) will inform strategic priorities, enable more targeted policy measures and permit more robust monitoring and evaluation of policy impacts (Poel et al., 2015). However, a recent survey by the OECD concludes that the public sector (outside the intelligence and security services) is in many ways lagging behind the private sector in the use of Big Data (OECD, 2015.

A case in point is that part of government which has an explicit remit to support 'ideas successfully applied', according to Dodgson et al.'s, (2013) parsimonious formulation of innovation. This is ironic, since innovation policy is one of the fields of government's activity that could benefit most greatly from the Big Data revolution, and where the limitations of traditional sources of data and market intelligence are most severe.

We believe that this will change. This paper draws on the innovation studies literature and case studies – including those based on our work at Nesta, an innovation foundation based in London – to argue that data analytics will in the future play a much more significant role in the making and monitoring of innovation policy. The paper is structured as follows:

In section 2, we overview a set of rationales and functions for innovation policy, while section 3 highlights shortcomings of traditional data sources and analytical techniques which limit their usefulness for accomplishing those functions. Section 4 discusses ways in which data analytics may address this, using illustrative case studies. In Section 5, we discuss the challenges that will need to be overcome, and the changes (including in organisation) that will need to be undertaken if data analytics is going to create value in innovation policy. Section 6 concludes.

2) Innovation policy: what, why and how?

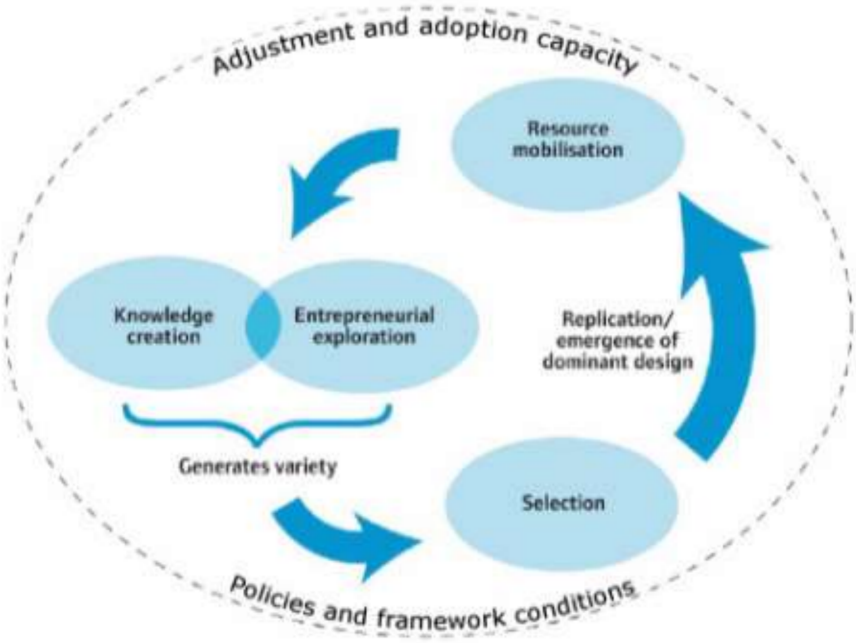
a) What?

If innovation is defined broadly as "*ideas, successfully applied in organizational outcomes and processes*" (Dodgson and Gann, 2013), innovation policy comprises "*all combined actions that are undertaken by public organizations that influence innovation processes*." (Borrás and Edquist, 2013). Innovation matters because it is widely understood to be the driver of long-term improvements in living standards (Aghion et al., 2009a) . It is also increasingly seen as the key to sustained improvements in social outcomes (Mulgan et al., 2007). Innovation policy might therefore reasonably be seen as the principal way that governments can influence the long-term wellbeing of citizens.

Innovation requires the cooperation of a variety of actors, including customers, large firms, SMEs, start-ups, universities, technology institutes, investors, intermediaries, consultants, sector organisations and employers, whose interactions are regulated by rules, or institutions (Klein Woolthuis et al., 2005). Innovation scholars conceptualise this complexity using models such as the Systems of Innovation framework (Nelson, 1993) and the Triple Helix of university-government-industry relations (Etzkowitz and Leydesdorff, 2000). Differences between places and sectors are captured with concepts like regional systems of innovation (Cooke et al., 1997) and sectoral systems of innovation (Malerba, 2002).

These approaches also stress the evolutionary processes governing innovation. Knowledge creators and entrepreneurs experiment with different ideas and technologies, and thereby generate variety. Experimentation is essential to manage the uncertainties that surround new (or novel combinations of) technologies and applications (Hekkert et al., 2007), and also market opportunities (Bakhshi, Freeman and Potts, 2011). Evaluation mechanisms select between 'successful' and 'unsuccessful' ideas; the former draw more resources, gain legitimacy (Aldrich and Fiol, 1994) , replicate and expand, while the latter ones fold. As a dominant design emerges, attention switches from exploring new alternatives to exploiting economies of scale. Figure 1 is a stylised depiction of the various functions of the innovation system (Johnson, 2001;(Bakhshi et al., 2009)).

Figure 1: A functional model of an innovation system



b) Why?

There are a number of theoretical reasons for thinking that governments can influence innovation processes to improve innovation outcomes. As with other areas of policy intervention, economists frame this in terms of "market failure" – in this case, linked to the fact that self-interested innovators typically fail to fully capture the benefits of their investments, with the implication that the level of investment in innovation will be lower than what is socially optimal (Gustafsson and Autio, 2011). If the social costs of this market failure are higher than the social costs of government intervention, the logic goes, there is a *prima facie* case for policy interventions that redistribute resources towards investment in innovation. In this treatment therefore, the innovation problem is one of allocative efficiency (Davidson and Potts, 2016).

Others argue that allocative efficiency is an inadequate guide for innovation policymaking, which should focus more on the dynamic implications of uncertainty and endemic coordination failure (Dopfer and Potts, 2008). Because the benefits of an innovation are uncertain until it is applied, firms may prefer to wait and learn from others before adopting it (Hall (2005) in Fagerberg et al., 2006). But if everyone acts in the same way, adoption will be delayed. This problem may also arise with technologies that generate spillovers into other industries, and with technological infrastructures like transport or telecommunications, where

investments made under uncertainty can generate "lock-in" to suboptimal solutions (Aghion et al., 2009b).

There are also instances where misalignment in incentives between different actors in the innovation system results in suboptimal outcomes (this is a perennial issue in the relationships between e.g. university-industry, investors-entrepreneurs, and between academic researchers operating in different disciplines). There may, for example, be insufficient knowledge exchange, interaction and cooperation between different groups (Klein Woolthuis et al., 2005). Innovation scholars use the term "systems failure" to describe such instances. Hutton and Schneider (2008) argue that because of uncertainty and systems failures the state has a role to play in demonstrating the benefits of new technologies, thereby encouraging innovation. Mazzucato, 2015 goes further, and argues that the private sector is often only willing to invest in innovation once the 'entrepreneurial' state has made the high-risk investments. Bakhshi, Freeman and Potts (2011) argue that the resolution of entrepreneurial uncertainty, through the creation and public sharing of knowledge about new economic opportunities and constraints should be the principal aim of innovation policy. For all these reasons, governments can help to address problems like "inhibited emergence", where a state of high uncertainty about the future configuration of a market or technology field hinders the emergence of new industries (Gustafsson and Autio, 2011).

In practice of course, as with other areas of public policy, the design and implementation of innovation policies reflects a range of motivations and considerations beyond theory (Flanagan and Uyarra, 2016). For example, experience suggests that how a policy is implemented depends critically on local context. Uyarra and Ramlogan (2016) give the example of cluster policies, where differences in policy outcomes can be attributed to context-specific institutional arrangements and policy path dependencies, not just variations in policy design and implementation. To take another example, if policymakers have bounded rationality or have limited access to the necessary data – not unreasonable assumptions in the high uncertainty environment in which innovation policy is set – they may not in fact be able to establish the 'correct' justification, market failure, system failure or otherwise (Cairney, 2016). The implication in both these cases is that policymakers may make better decisions if they have access to more detailed, relevant information and have more sophisticated capabilities to make use of it.

c) How?

Many different policy instruments have been used to remove the potential bottlenecks to innovation. Edler et al., (2016) provide a typology of innovation policies and present meta-evaluations for 16 sets of innovation policy instruments, including activities as diverse as

fiscal incentives for R&D, skills policies, cluster policies, entrepreneurship policies, networking initiatives, public procurement and technology foresight.
Edler et al. define innovation policy broadly as "*public intervention to support the generation and diffusion of innovation, whereby an innovation is a new product, service, process or business model that is to be put to use, commercially or non-commercially*." As such, we would add to their list of innovation policies:

- The act of official classification and measurement of innovative technology applications and sectors, as by legitimising such activities in the eyes of innovation systems actors, statistics produced by government can be a potent form of policy support (Tech City UK/Nesta, 2016), and
- Open data, whereby growing numbers of governments hope to boost enterprise and innovation through making "*digital, machine-readable data available to business and citizens to use and reuse free of charge*" (World Bank, 2014).

Edler et al. also note that the locus of the design and implementation of innovation policy varies within government. In particular, it is on the one hand the responsibility of different tiers of government, including central government, cities and local authorities, and it cuts across different functional ministries (like health, transport as well as the more obvious business, science or innovation) on the other. The implication for the current paper is clear: in considering what are the opportunities from data analytics for innovation policy we must think beyond a narrow focus on the data needs of business ministries.

3) Limitations of existing data sources

Of course, governments need data to inform the design and evaluation of innovation policy in all its guises – including the publication of official statistics and open data. In this sense, innovation policy is no different to other forms of public policy, and has the same requirements for data as other "evidence-informed decision-making" (Nesta, 2016): policies need to be prioritised, targeted, designed, implemented, evaluated and adapted (or wound down). Data has a role to play in all stages of that policy cycle.

In general, innovation policymakers have tended to rely on the same methods for collecting data as other policymakers – such as survey instruments, which ask firms about their various innovation activities (Frascati, Oslo) and company accounts, which contain measures of their financial performance (Amadeus, Orbis). Where administrative data –information collected primarily for administrative (not research) purposes – has been used, it has tended to relate to the most 'visible' parts of the innovation system, such as investment in R&D and patent

volumes. However, what is most visible is not necessarily the most important for innovation outcomes, nor where policy necessarily can have its greatest impact.

Even in cases where analysts have attempted to collect data on more 'hidden' aspects of innovation, such as user-led innovation (Von Hippel et al., 2012), collaborative innovation (Tether, 2002; Tether and Tajar, 2008), knowledge exchange (Hughes and Kitson, 2012), services innovation (Abreu et al., 2010) and design (Galindo-Rueda and Millot, 2015), there has been a tendency to employ survey approaches.

In the case of survey data collected by governments or national statistical institutes and official administrative data, the value for timely policymaking is further constrained by the lags associated with bureaucracy and, in the case of administrative data, the time required to clean and structure the data in a way that makes it useful for analytical purposes. It is not uncommon, for example, to see innovation policy strategies in OECD countries that use data that is several years out of date. One area where there is room for timely administrative data collection is in the area of programmatic interventions, though published examples are rare. Potts et al., 2016 discuss how the officials at the Department of Economic Development, Jobs, Transport and Resources in the Victoria State Government collect primary data on market challenges and solutions from technology firms they are supporting for the purposes of identifying businesses working on similar challenges, thereby informing the scope of their interventions.

In parallel with the use of standard techniques for collecting quantitative data, notwithstanding important contributions using qualitative and case study methods (Freeman, 1982, 1987, see also OECD Sectoral Case Studies in Innovation), innovation policy analysts have also tended to use the same analytical techniques as those researching other policy domains. In particular, econometric models have been used to try and establish causal relationships between innovation behaviours and performance at the firm level and interventions and outcomes at the policy level (Encaoua et al., 2013). However, that innovation reflects the complex interactions of large numbers of institutions, as discussed in Section 2 of this paper, and is influenced by unpredictable external influences creates particular challenges for identification (Bakhshi et al., 2013).

Although the relational nature of innovation is acknowledged by theory as discussed in section 2, operationalizing it in a quantitative setting and introducing it into policymaking has been harder, in part because surveys can only capture crudely, if at all, the interactions between a respondent and others actors. An exception to this is when knowledge creators publish academic papers and/or patents: in those cases, citation patterns can be used to

reconstruct knowledge networks, and to estimate the influence or impact of an innovation (Hood and Wilson, 2001). This kind of information is used extensively to produce metrics for international benchmarking of countries' innovation performance, and for evaluating the impact of publicly funded research (Smith et al., 2011). There is less evidence of uses of this data in a 'network format', or to inform decision-making earlier in the policy cycle (e.g. when designing or targeting interventions to address innovation system failures). None of these things are possible where innovators do not publish academic papers or patents.

Another limitation in official data is that it can only be queried using standardised industrial and occupational classifications. These internationally agreed standards cannot, by definition, capture industries (occupations) that did not exist before they were agreed, which makes them unsuitable for monitoring the development of new industries (occupations), and limits the government's ability to play a role in removing barriers that might inhibit their emergence – which, as we have discussed, is an important rationale for innovation policy.

From a policy design and delivery perspective, being able to identify stakeholders in new sectors and technology areas to identify their needs (potentially different from those in existing sectors) is another activity hard to undertake using existing data sources based on standardised codes – this policy activity is further hindered by the fact that most official data is only available in an anonymised format to avoid disclosure. As before, patents and publications which contain both standardised metadata (such as patent classes and keywords), unstructured textual descriptions and relational information lends themselves better to the analysis of the emergence of novelty and the identification of specific innovative organisations (Aharonson and Schilling, 2016; Breitzman and Thomas, 2015; Mathew, 2015). However, they are made available only with a significant lag, and as before, are only suitable for the study of new sectors that patent or publish academic research.

4) Opportunities and state of the art in innovation data analytics

The scale and rapidity of the data revolution is well rehearsed (Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013; OECD, 2014): the digitisation of business processes, commerce and social life has brought with it an explosion of data, often generated as a by-product of other activities – in common with administrative data. Private firms create websites to promote their products and services, and the content of these websites can be analysed to identify their industry; individuals connect with each other in social media, generating evidence about the structure of social and industrial networks. The increasing availability of government data – either publicly, as with 'open data', or restricted access via secure facilities – is an important part of this phenomenon (Ubaldi, 2013), as is the

proliferation of technologies, services, tools and methods to collect, store and process, analyse and present data (see Table 1).

There are several ways in which some of these new datasets and methods can help address the data needs of innovation policymakers, overcoming some of the limitations of existing datasets that we identified earlier. We review them now, together with relevant work and, at the end of each subsection, a more detailed case study (see table 2 for a summary).

Table 1: Summary of data technologies, tools and methods

Function	Examples
Data collection	• Web crawling and scraping software; • Application Programming Interfaces (APIs) to access data automatically from websites and online databases.
Data storage and processing	• Database technologies, including 'non-relational' databases to store non-tabular data efficiently, • Cloud services for remote storage and parallel processing of data; • 'Big data' framework to distribute data jobs across many machines, and speed up analysis.
Data analysis	• Software applications for data analysis; supervised and unsupervised machine learning methods for classification, prediction and clustering, natural language processing and social network analysis.
Data reporting and visualisation	• Tools and products for interactive data visualisation, reports, dashboards and data-driven websites.

a) Unstructured data can be used to identify emerging areas of technological and economic activity in a more accurate, comprehensive and timely way

By contrast with official data, which is structured around standardised industrial and occupational codes, most of the dataset types we cinclude in Table 1 are unstructured or semi-unstructured. For example, a company website contains information about its activities in textual format. It is possible to analyse this text in order to identify entities of interest (for example, businesses operating in a new sector), or define new categories of activity 'from the bottom up'.

This data is also more likely to be updated frequently than company SIC codes, meaning that its analysis has the potential to yield timely result

Unstructured data can be analysed in a variety of ways, ranging from relatively simple strategies based on a search for keywords of interest, to unsupervised machine learning methods and text mining.

An example of the former approach can be found in Gök et al., (2014), where the authors use a structured keyword search in commercial business databases to identify "Green product" businesses with no representation in the Standard Industrial Classification (SIC) codes. Shapira et al., 2010 use a similar strategy to study the emergence and evolution of the nanotechnology industry. Mateos-Garcia and Gardiner (2016) perform a similar analysis of 'tech meetup' activity in the UK, finding that community interest on emerging tech topics tends to respond to events outside, such as company acquisitions, product launches and innovation breakthroughs .

One downside of keyword-base searches is that they rely on the comprehensiveness of the initial taxonomy created by researchers or domain experts – this approach will not return records labelled with 'synonyms', and by definition, it is not suitable to identify completely new topics the researcher is not aware of before commencing the analysis. Topic extraction methods developed in the field of Natural Language Processing make it possible to address some of these issues, by drawing on the co-occurrence of words inside documents (e.g. a company website) to identify interrelated topics (Blei, 2012). Although there has been an explosion in the literature using such methods on online data, most of them are focused on public debates in social media, and are perhaps of less direct interest for innovation policymakers.

An exception is Nathan and Rosso, 2015, who use a business dataset labelled with industry keywords generated via machine learning methods in order to analyse the size and evolution of the Information and Communication Technology (ICT) industries in the UK. Their findings suggest that official data significantly underestimates the size of the digital tech industries, in part because it fails to capture businesses classified outside of the ICT-related SIC codes but which nonetheless make intensive use of digital technologies – think of a financial services business using High-Frequency Trading algorithms, for example, or an advertising company specialising in Search Engine Optimisation.

Case study: Mapping the UK video games industry using 'web data'.

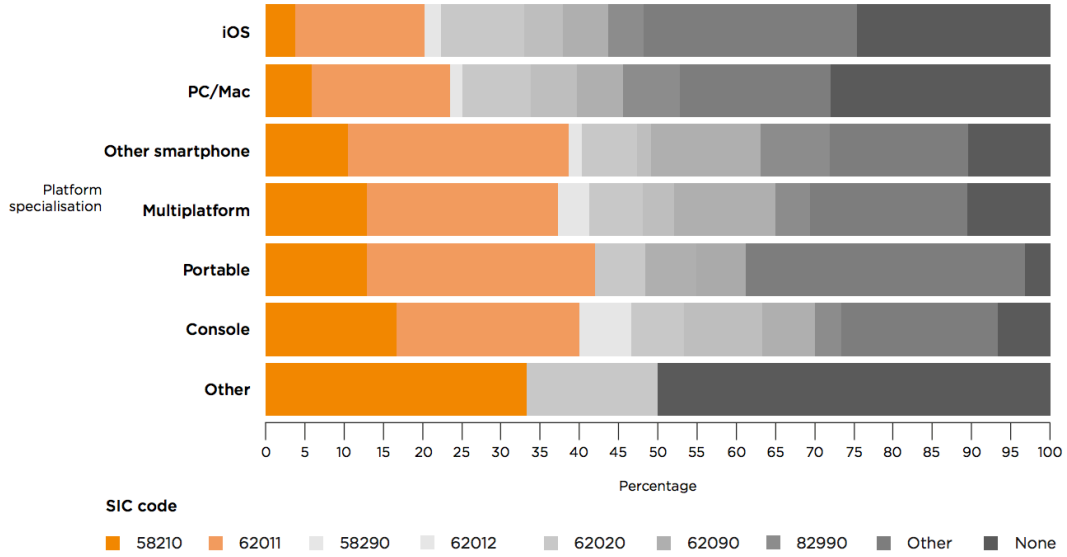
The video games industry provides an interesting case study of the challenges involved in measuring young, fast-growing, innovative industries, and of the opportunities presented by new data sources.

In recent years, there has been a raft of government policies put in place to support the UK video games industry. There is, however, a general recognition that video games are inadequately captured in official government statistics, and this diminishes the effectiveness of these policies. Until 2007, the industry did not have its own SIC code, and it was therefore not possible at all to measure its size and evolution using official data. In the past, researchers have created data about the sector through surveys targeting samples drawn from company lists compiled by industry trade bodies and specialist consultants. This approach is expensive to administer, and faces the risk of potential biases in its sample design.

Mateos-Garcia, Bakhshi and Lenel, 2014, adopt an alternative approach, using product directories, review sites and digital distribution platforms with detailed information about video games products and their producers to identify companies in the sector. After scraping data about 73,148 games companies globally, they match it with the UK business register and perform additional manual validation of a sample and model the results using decision trees to predict the status of non-manually validated companies. This results in a list of 1,902 active games companies in the UK, together with their specialisation profile (games platforms that they target), addresses, SIC codes and (for large companies) financial information.

The analysis of this dataset confirms widespread concerns about misclassification andmis-measurement of UK video games businesses: only a third of the companies in the dataset are in the official SIC codes. Misclassification seems to be more severe with younger companies, and companies operating in new platforms for games distribution, such as smartphones and tablets. This is consistent with the idea that innovative businesses in particular are poorly captured by official video games SIC codes (see chart below, where the orange bars represent SIC codes in the official definition of the sector).

Figure 1: SIC code distribution by games platform



The authors also benchmark maps of the video games industry using web data with results based on business register data and official SIC codes. Although this comparison reveals a high degree of consistency in the places characterised by high agglomerations of games businesses according to both approaches, it also shows that official data misses a long tail of UK locations with some games-production activity – information which is of interest to local and national policymakers.

This approach also makes it possible to: identify companies that are economically active yet too young to have selected a SIC code in the business register; access data which is relevant for the sector but is not collected through official business surveys, and provide a higher level of resolution (company level) than is possible with official sources (for reasons of data disclosure), enabling some of the open data analysis and strategies we refer to later in the section.

This kind of analysis is not without its challenges for researchers, however. Quality assurance is critical, as the data will not have undergone the national statistics institute's quality procedures. The automated, fuzzy matching and classification of companies used in this research is probabilistic in nature, and a source of false positives/false negatives that are easy to spot in a highly transparent dataset. Manually removing such classification errors has high fixed costs, yet is important in order to maintain the policy credibility of the data set. The importance of combining automated data-collection processes and domain knowledge from industry experts in order to strike the right balance between scalability and reliability is an issue we return to in section 5.

b) Relational data can be used to map networks of innovative activity and collaboration

The web is by definition a networked (hyperlinked) system, and many of its most popular platforms – such as Google, Wikipedia, Facebook or Twitter – leverage that structure to offer relevant information to users, and encourage them to interact and trade (Brin and Page, 2012; O'Reilly, 2007). The data generated by all this activity can be used to reconstruct social and industrial networks that in some cases may be of interest for innovation policymakers, particularly regarding industries whose networking patterns are hard to study through publication or patenting networks.

Tambe, 2013, for example, analyses labour mobility of engineers with big data skills between tech businesses in the USA using LinkedIn data, and this way, measures how the skills investments undertaken by some businesses spill into the industrial clusters around them. Mateos-Garcia and Bakhshi (2016) use data about participation in Meetups in order to measure absolute levels of networking in creative industries clusters in the UK, as well as networking between different communities locally, regionally and internationally. Their analysis reveals different patterns of networking across clusters: dense, urban spaces have high levels of networking, often involving crossover between communities in different sectors. By contrast, creative clusters in suburban areas do less networking, and what they do is confined to their own industries.

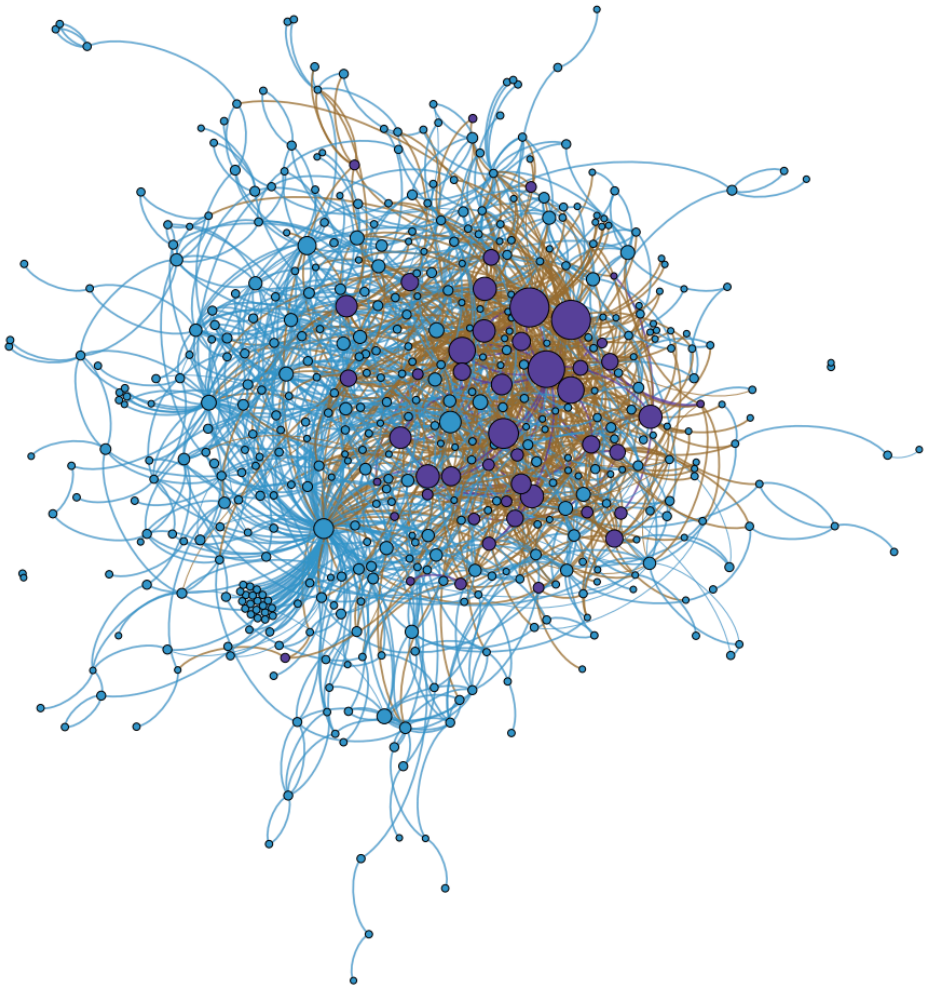
Case study: Using social media data to measure the impact of events

Social media data also presents opportunities to measure the impact of events and conferences. Such activities, often publicly supported with the goal of strengthening innovation systems, are rarely evaluated in a way that considers their impact on the structure of those systems, or their additionality compared to networking activities that would have taken place without the intervention anyway (e.g. because attendees are already 'close' to each other geographically or socially) (Cunningham and Ramlogan, 2012; Giuliani and Pietrobelli, 2011). This is partly for lack of data: collecting detailed data about networking at events using tools such as surveys can be prohibitively costly when there are a large number of participants. Moreover, dependence between observations and high skewness in the degree (link) distribution of networks makes it difficult to infer impacts from random samples.

Bakhshi, Mateos-Garcia and Davies (2014) seek to address some of these issues using Twitter data to measure the formation of new connections and communication exchanges amongst attendees at a tech conference. In order to do this, they scrape the Twitter ids of 702 attendees from the conference website (around half of the total), and reconstruct their network before

and after the event (the chart below displays the new follow connections formed between attendees at the event – the purple dots represent speakers).

Figure 2: New Twitter follow connections at a tech conference



Their analysis suggests that the event spurs high levels of connectivity between participants. During the time of the conference, attendees connected with each other in Twitter at a rate 4.6 times faster than what they were doing with Twitter users outside. Network additionality was explored by measuring the distance between attendees in the twitter network before the event. The analysis shows that although many of the connections formed at the event involved individuals only one step removed (this was the case with over a quarter of the new 'reciprocal' connections), a significant number of connections (20%) involved people who were not at all connected in the network before the event – this was especially the case where attendees had travelled to the event from different countries, and therefore intuitively were less likely to have overlapping networks. The analysis of communication between attendees at twitter reveals higher rates of information exchange between newly formed connections

than between existing ones, consistent with the idea that these new connections opened up access to new and potentially valuable information.

There are some important limitations for this analysis, however: social media data often lacks important metadata which is necessary in order to interpret the findings. For example, not all event attendees provided information about their location in their Twitter profiles, so this information had to be gathered manually in order to enable the comparison between social and geographical distances mentioned above. Perhaps more significantly, the measures of impact (connections formed and communication flows) are weak proxies for the 'practical' and economic impacts that innovation policymakers are interested in.

In follow-up work, Cronin et al (2015) try to address this issue by matching Twitter and GitHub data via user ids to analyse whether participation in tech conferences is associated to improved outcomes in software projects. Their analysis shows that GitHub developers who were central to the network of participants in multiple tech events consistently attracted much greater code contributions to their projects than GitHub users in the control group. In the final year of code contributions examined, 2013, event participants had 112 per cent more contributions to their GitHub projects than the random GitHub users. While this analysis is not without limitations – in particular the risk of reverse causality, if successful software developers have more incentives to attend tech events in order to promote their projects – it illustrates how creative combinations of data sets can help start measuring the impact of as-yet poorly evidenced innovation interventions.

c) Linked and linkable datasets can offer a better understanding of complex innovation systems and dynamics

If the data characterising complex innovation ecosystems and processes is fragmented across many sources, then there is the risk of obtaining a partial view of the situation, and incomplete information for policy. This is an important concern when developing policies to support industrial clusters reliant on a variety of inputs and social, physical and digital infrastructures and where local institutional context matters, or when trying to understand the impact of innovation interventions whose value may be captured not by a company but by its employees, or by other businesses in its value chain, and realised over long periods of time – such as the innovation networks we just discussed.

The concept of alt.metrics for scientific research is a well-known example of how existing data about publications can be combined with new social media data in order to measure the diffusion of scientific research outside academe, improving our understanding of the public

value of research, as well as generating metrics of interest on research more timely than citations accruing over years (Piwowar, 2013; Priem et al., 2010).

The growing availability of data in linked or linkable formats, both as 'open data' that can be accessed by anyone, and administrative micro-data in secure environments, is also helping us to understand the micro-dynamics of labour markets and business growth. There are many recent examples showing how these big – if not necessarily new, or unstructured datasets – can be used to shed light on economics issues of great interest for innovation policymakers (Einav and Levin, 2013; Horton and Tambe, 2015), or to evaluate direct support for innovation (Department for Business, Innovation and Skills, 2014). The movement towards collecting and using transactional data that is relevant for innovation is not confined to tax records and company financials. The UMETRICS programme captures administrative information about federally funded researchers in a consortium of 15 US universities, thus generating data to guide public investments in science, and to evaluate its impacts (Lane et al., 2014). In the UK, the Gateway to Research platform provides an open API into a comprehensive dataset about all Research Council-funded projects in the UK. This data is already being used to map research networks in different disciplines with the goal of identifying new opportunities for collaboration between universities and businesses (Mateos-Garcia and Gardiner, 2016).

Case study: Combining official, open and web data to map digital ecosystems in *Tech Nation 2016*

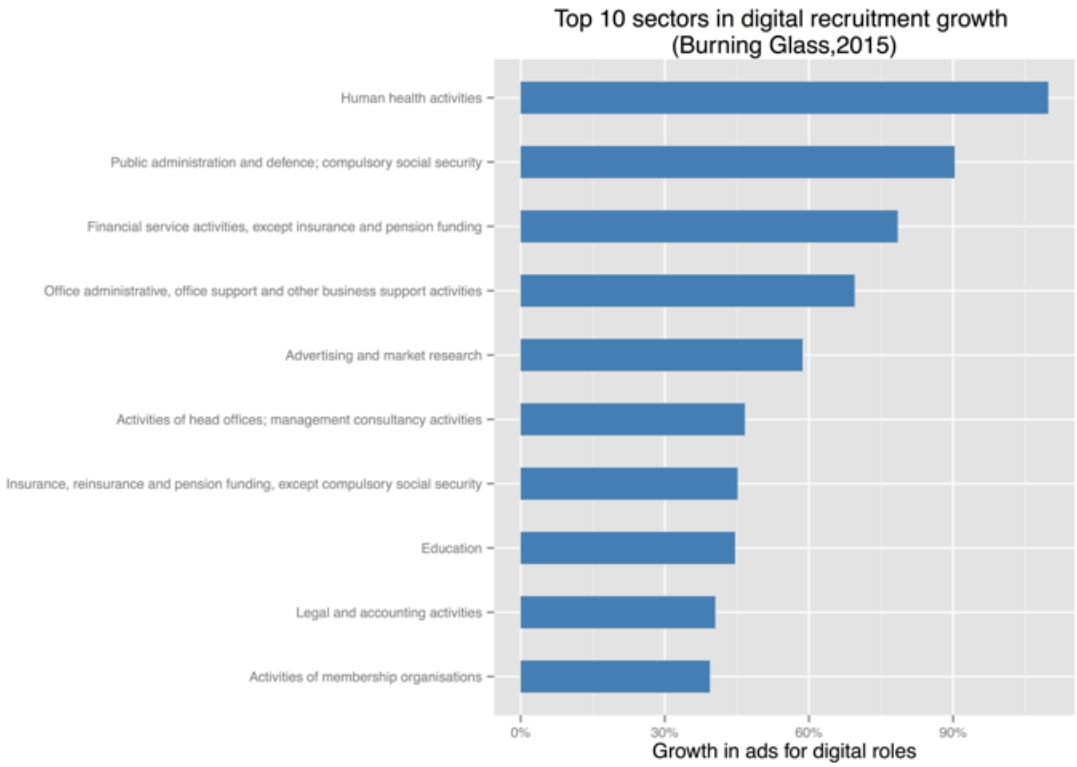
Going beyond administrative datasets, other projects are combining official, open and web data to generate comprehensive profiles of innovation ecosystems. Some recent examples include the World Bank's analysis of digital tech start-ups in cities (Mulas et al., 2015), and Nesta reports looking at digital technology TechCity UK/Nesta, 2016) and creative industries clusters (Mateos-Garcia and Bakhshi, 2016). In all these cases, there is a strong emphasis on measuring the scale and evolution of economic activity, and complementing this data with information about local resources and capabilities that could act as drivers of – or barriers to – growth, often based on novel data sources and data combinations.

In the example of *Tech Nation 2016*, this involves official business and labour surveys and SIC codes to generate estimates of economic activity (employment, turnover and GVA) in digital tech clusters in the UK, as well as their evolution, together with other web-derived data shedding light on important, but poorly understood, features of the digital tech industries. This includes an analysis of content scraped from company websites in order to characterise their sector of operation at a finer level of resolution than is possible with SIC codes, and an

analysis of 6.5 million online jobs ads to estimate demand for 'digital roles' in different locations and sectors, as well as the salaries compared with other roles.

Leaving aside headline statistics of interest to national and local policymakers, this report highlights the extent to which digitisation is not a phenomenon confined to a small number of sectors, but pervading the economy in a way that is difficult to understand using 'mutually exclusive, collectively exhaustive' industrial categories such as SIC codes (for an illustration of this using job ad data, see Figure 4).

Figure 3: Digital recruitment in non-digital sectors (Tech Nation 2016)



One lesson from this work is the importance of triangulating the findings coming from different datasets, some of which are new and as yet untested. In the case of *Tech Nation 2016*, this went beyond the need to reconcile the findings of the analysis based on web data, which is arguably more timely and granular but lacks important economic variables, and may suffer from selection biases (i.e. variation in business propensity to create a website, and variation in website quality allowing a valid analysis using topic modelling methods) with the results of official data. In one instance, it was also necessary to reconcile the findings of ostensibly similar analyses of company activity using web data which were undertaken by different organisations, with different methods and different results. In the conclusion, we

will come back to the risks that a real (or perceived) lack of robustness could create for the application of new data sources in innovation policy, and potential remedial strategies.

d) Open data and open analysis can lower barriers to further analysis and application of data

We just highlighted how open data can act as an input into policy-relevant researches and analyses. However, that is not its only potential application. We know that there is a growing market for solutions that help workers, businesses and investors find jobs, partners and prospects, including LinkedIn, GlassDoor, DueDil, CrunchBase, or Angel List to name but a few. There is no reason why these same actors could not benefit from some of the datasets and analyses we mentioned above to make their innovation activities more effective. In this scenario, innovation data analytics would impact innovation indirectly, by improving the quality of innovation policymaking, and directly, by reducing uncertainty about the market performance of different innovations, and lowering the costs of finding partners and collaborators. This is in line with the idea that open data can 'unlock significant economic value' (Manyika et al., n.d.).

A perhaps subtler but also significant, potential benefit from greater availability of open innovation data for further processing, linking and analysis, may be to open up the processes through which Science, Technology and Innovation indicators are constructed, therefore making them more 'socially robust', and supporting collective learning during their development (Barré, 2010; Stirling, 2014)

In order to realise all of this value, however, it is important to invest in the quality of the data being re-distributed, and also in its accessibility for users who will not always be technically proficient (Hidalgo, n.d.). In recent years, there have been several efforts in this space which utilise software innovations in web development and design. This includes a raft of cluster maps using open and web data (http://www.techcitymap.com/index.html, http://www.camclustermap.com/) and interactive data visualisations of the economy of Brazil (http://dataviva.info/en/) and the USA (http://datausa.io/). The hypothesis that it is possible to use network science to map industrial networks along different dimensions (geographical, cognitive, social), and use that information to help businesses decide whom to collaborate with is being tested in the Netsights project in Denmark (http://www.netsights.dk/).

Finally, there has been an explosion of interactive data visualisations in the fields of journalism and design, with a focus on telling compelling stories with data. We are starting to see an increasing number of visualisations in the domain of innovation. Some examples from Nesta include an analysis of antibiotic resistance in the EU (part of an open challenge prize to

develop solutions to this important issue: http://www.nesta.org.uk/blog/fight-against-antimicrobial-resistance-across-europe), and a visualisation of trends in the demand for skills using data scraped from online job boards (http://www.nesta.org.uk/blog/top-30-skills-chart).

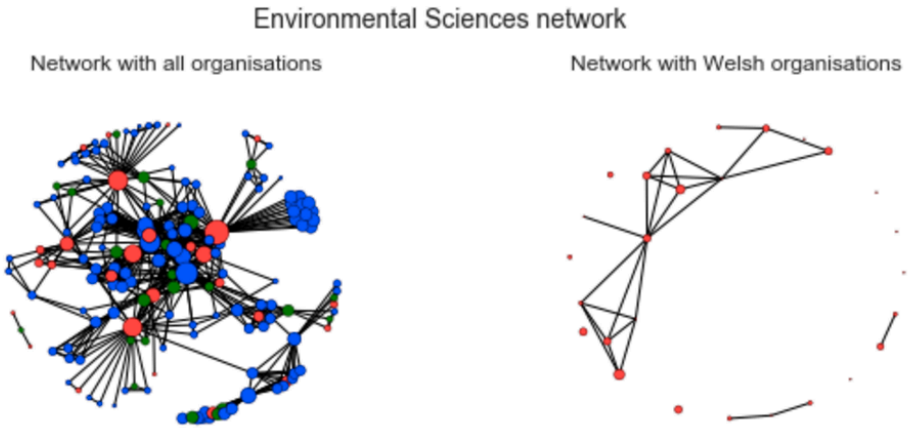
Case study: Arloesiadur, building an innovation data analytics platform for Welsh Government

Our final case study concerns an ongoing research collaboration between Nesta and the Welsh Government to build a data analytics platform which will be used by innovation policymakers in Wales. When it is launched in 2017, the platform will combine data coming from web, open and official data sources, and automate their analysis and visualisation in order to offer users in different departments with relevant information supporting decision-making across the policy cycle.

The ultimate vision for the pilot is to create evidence about the value of big data for innovation policy in all the areas discussed in this section, including the publication of open data, data visualisations and code to supports follow-on analyses and decisions, also outside of government.

Although the project only started in January 2016, there are already some interesting lessons emerging from it. One theme coming out from the scoping of the platform with users in government is that they do not see innovation analytics as a tool to automate decision-making, but as another source of potentially useful information, and even the starting point for other research efforts using qualitative and "traditional methods" such as surveys, which can generate policy-relevant outputs harder to obtain from web and big data-sets (e.g. understanding barriers to innovation, generating case studies of policy impacts and exemplars of innovation and entrepreneurship). The use cases for the data are also turning out to be wider than initially expected. In particular, there is substantial interest on the potential use of the outputs from innovation data analytics in the operational delivery of innovation policies, and not just for strategic decision-making and evaluation. One use case for this is helping innovation policymakers identify businesses and universities currently participating in research projects with partners elsewhere in the UK but not in Wales, who may be encouraged to also engage in local collaboration networks (the Figure below, based on an analysis of open government data about research-council funded projects, displays the research collaboration network of Welsh organisations around the discipline of Environmental Science before and after removing partners outside of Wales; the isolated nodes and components in the right panel could hypothetically be connected with each other to make the national network denser).

Figure 4: Environmental Science research networks involving Welsh organisations



Options for redistributing open data may be constrained by the terms and conditions for the usage of data extracted from commercial platforms, as well as restrictions in the re-release of personal data. For example, the network above could be plotted using individuals as nodes, instead of organisations, but there are ethical issues in the release of data at that level of resolution. Striking the right balance between the privacy rights of individuals included in innovation datasets, the commercial interests of web platforms who are collecting some of this data, and the data needs of innovation policymakers is an important issue we touch on next section.

5) Challenges

The literature review and case studies in the previous section focused on the big data and data analytics opportunities for innovation policymakers. In this section, we turn to the challenges that need to be overcome in order to realise those opportunities, organised around the logic of a 'data pipeline' where data is first collected, then analysed, and finally used to generate an impact.

a) Quality

If new data sources are going to play a role in innovation policy, then they need to be of sufficiently high quality – but what are those dimensions of quality? Two important dimensions are representation (does the data capture a population of interest, or is it biased?) and accuracy (do the metrics represent the 'real-world' behaviours or activities of interest, and is this representation consistent over time?).

It is now clear that initial assessments of these two dimensions of quality in web and big data were excessively optimistic. To begin with, and contrary to claims that with big data N (the sample size) equals "All", therefore removing response biases in surveys, as well as the need to infer statistically population attributes from samples (Anderson, 2008; Mayer-Schönberger and Cukier, 2013), the reality is that participation in web platforms and services does have important biases. For example, a higher usage of Twitter among young, affluent segments of the population limits its usefulness as a gauge for public opinion, and as an input into policymaking, for example during public emergencies (this was starkly manifested in the vast underrepresentation of less affluent neighbourhoods in New York from maps about the impact of Hurricane Sandy based on Twitter data, (Crawford, 2013; Crawford and others, 2011)).

Data accuracy is also a substantial concern in a world where digital platforms are often hacked, or used to distribute spam. Even where they are not, changes in the behaviours of their users can introduce breaks in time series that are unrelated to the activities that analysts want to measure. A famous example of this is the precipitous decline in the accuracy of Google's Flu Trends service (which predicted the incidence of the flu using web search data) as a consequence of changes in the behaviours of web users, apparently linked to shifts in the reporting of flu outbreaks in the media (Lazer et al., 2014).

These biases also have implications for innovation data analytics. For example, Twitter is used as a professional networking tool in digital media and the creative industries, and therefore its levels of adoption in that sector could be expected to be higher than, say, in biotechnology. This means that analyses of biotechnology in twitter are likely to capture unrepresentative segments of that sector and an skewed view of it. Similarly, if communities of technologists display geographical or sectoral variation in their propensity to use Meetup.com as a platform for networking versus other websites, or simply prefer to arrange their events via e-mail, then this may reduce the external validity of any analyses using data from that platform. The same contention applies to other websites and services.

Data accuracy is a further issue. As an illustration, in the analysis of Twitter data to measure the impact of events that we summarised earlier, Bakhshi et al (2014) identified a single attendee who had been following 556 event attendees before the event (almost 80 per cent of the participants for which there was data), and un-followed 485 of them afterwards. This individual outlier was, singlehandedly, changing some of the report's results. This illustrates some of the challenges of working with social media data – as well as the need for caution in collecting and analysing it.

What to do about all this?

In general, it seems that many of the new web sources discussed above are particularly suitable for the analysis of digital tech industries that display particular innovation behaviours (high levels of face-to-face networking and peer-based training, and participation in online communities) and produce certain types of outputs (software, creative consumer goods). This is not a reason not to use these sources for more general analysis (in the same way in which we do not stop using patents or citations even though these metrics are only relevant for a small number of sectors, or present biases due to self-citing, strategic patenting etc.), but a reason to be careful about their limitations and, as far as possible, draw on domain expertise (for example, knowledge of the industry or ecosystem being analysed) when scoping data sources, and interpreting their findings. As previously mentioned, triangulating between different data sources is a good strategy for identifying 'puzzles' and quirks in the data that may raise concerns about its quality (Bean, 2016). In order to encourage 'collective learning' in this space, it would be desirable for innovation data analysts, data scientists and policy users to as much as possible make their analytical inputs and outputs (data and code) publicly available in order to develop a shared understanding of the limitations and valid use-cases for different datasets, including their relevance for different countries (an issue that needs to be addressed in order to achieve international comparability). It seems that innovation data analytics needs its own push for reproducibility we are seeing in other fields (Collaboration, 2015; Peng, 2011) .

It is also worth bearing in mind that some of the issues discussed above are particularly problematic when analysts and policy users are interested in generalising from a sample but less so, in other use cases – for example, if the analysis is going to be used to identify emerging trends, or provide an early warning sign of developments that would otherwise remain undetected. Similarly, policymakers using innovation data analytics for operational purposes and targeting (e.g. to identify specific businesses to interact with while scoping a new programme) may be less concerned about generalisation specially where these methods are generating new 'leads' that they did not have access to before, or if they are looking for 'outlying' actors such as high-growth firms, high-performing entrepreneurs etc.

b) Application

Even the best quality data will not be applied if its analysis does not yield actionable information, or if this information is not trusted by policymakers. On the first issue, while the timeliness of new data sources can help analysts respond faster to policy needs, lack of policy-relevant metadata (e.g. financial performance in businesses identified online, or tangible outcomes linked to innovation activities such as networking), and contextual information helping to interpret the meaning of any given pattern, can reduce its usefulness.

For instance, previously we mentioned that we had found highly specialised networking in some UK creative clusters. Even assuming that this reflects an economic reality, it is not clear what a policymaker should do about it: Is it a problem? What are its causes? What are the solutions? Domain expertise (e.g. an understanding of the academic literature regarding the impact of different networking structures on cluster resilience) and local knowledge can help make sense of patterns in web data, and decide what the policy response should be (if needed). The fact that many of the datasets we refer to in this report are public, and therefore allow identification of businesses, communities and networks for additional data collection and snow-balling make them more suitable for this iterative mode of innovation policymaking than official, anonymised statistics that some-times resemble black boxes.

This is not to say that big and web datasets do not have their own 'black box' issues, which are detrimental for their credibility (Einav and Levin, 2013; Pasquale, 2015). This is a particular issue where the analysis is based on proprietary data or algorithms that cannot be reproduced by other researchers – several of the studies discussed above involved data accessed under those conditions. This lack of transparency makes it difficult to analyse the quality of the data (i.e., whether it is representative or accurate), potentially leading to mistrust about the findings. Relatedly there is the risk that different analysts measuring the same variable will reach clashing conclusions because, for example, they use different algorithms and/or parameters. Without transparency about how the data was collected and analysed, it is difficult to judge the relative merits of different approaches, and to build trust on these new methods. In the field of scientometrics, researchers have started raising concerns about the proliferation of unstable science alt.metrics ((Thelwall et al., 2013)

Industrial classifications are another area where there are challenges: as we discussed in the previous section, one can generate a myriad industrial taxonomies using web and big data– the question, if we want to undertake international and historical comparisons, is which one to choose, and how to maintain its consistency over time. Here is where internationally agreed standards such as the SIC and SOC codes enjoy an advantage over bottom-up classifications linked to their relative rigidity, and the formal processes that regulate their update. If big and web data-based methods are going to generate useful and usable industrial classifications addressing the limitations of existing codes, a similar process of standardisation might be required. A greater degree of reproducibility and transparency around industrial classification efforts currently underway would be a first, valuable step in that direction.

Another potential 'black box' dimension of innovation data analytics pertains its reliance on complex machine learning models that in some instances – such as random forests or neural nets – have low levels of interpretability. The significant ethical challenges that this poses in

innovation policy are arguably less severe than other areas of policy such as policing or education (Rudin and Wagstaff, 2013, Armstrong 2015). Having said that, it is improbable that innovation policymakers will desire to act upon the recommendations of predictions generated by pure black box algorithms (e.g. 'what scientist to sponsor, what business to support, what technology to invest on") unless those are shown to deliver better outcomes. There is significant scope for experimentation in this area.

c) Organisation

Data analytics is like any other new technology, in that its successful adoption by an organisation requires complementary investments in skills, organisational processes and culture. We overview these in turn, in relation to innovation policy:

Firstly, much has been made of the need for new skills, and even a new role, the vaunted data scientist (Conway, 2010; Patil, 2011) for collecting, managing and analysing big data; there is also evidence of a shortage of professionals with this profile (Manyika et al., 2011; Mateos-Garcia et al., 2014). Innovation agencies seeking to develop an analytics capability will need to build teams through a combination of training and recruitment, locate those in the part of the organisation where they are able to create the biggest value, and manage projects in a way that encourages experimentation while retaining relevance for the rest of the organisation (Bakhshi et al., 2014b). In doing this, they can leverage a growing range of open source tools, online communities and open courses and datasets (Hardin et al., 2015).

Secondly, data analytics has the potential to transform the dissemination of information for decision-making inside innovation agencies and support bodies, and this has implications for how they are organised, and their processes. Although there has been a tendency to think of data analytics as a tool for automation and centralisation of decision-making (Medina, 2011), evidence from the private sector suggests that decentralisation and employee empowerment are stronger complements to the adoption of data-driven decision-making (Bakhshi et al., 2014a; Brynjolfsson et al., 2011). This is in line with the agenda of open data and analysis, and the synergies between local and domain knowledge and data analytics we have referred to throughout this paper.

Finally, it is clear that without a culture of respect for data and evidence, linked to the idea of evidence-informed policy, it is unlikely that novel and useful insights from data will be applied unless they support existing intuitions or political considerations (in which case their additional impact will be limited). Although finding the right balance between data implications, political considerations and managerial intuitions is as much of a challenge in innovation policy as in other policy domains (in some respects even more so, given high

levels of uncertainty about the future pervading the area of innovation), it is to be hoped that injections of data restrict the scope of undesirably discretionary decision-making, improve operational efficiency and support the evaluation of interventions and processes in a way conducive to finding, over time, and after much experimentation, a satisfactory balance.

6) Conclusions

This paper draws on established literatures and new experiments to explore how new data sources and analytical methods – data analytics – can address the needs of innovation policymakers, help them fulfil their functions, and support the successful application of new ideas. The picture that emerges from our assessment suggests that, at least for now, the role of data analytics in policy will be incremental rather than transformative, complementing existing datasets, decision-making processes and domain knowledge bases, and improving the effectiveness of policy operations rather than leading to a radical overhaul. We posit two complementary reasons for this. First, the volume of 'transactions' (decisions that could be based on data) in innovation policy is lower than in other policy domains such as Social Policy, Education or Policing: the scope for automating decision-making is therefore lower. Second, innovation policy decisions are frequently made in a context of high uncertainty for which predictive models based on historical data are less suitable. We would as a consequence expect intuition, imagination and human judgement to continue playing important roles in innovation policy – perhaps to the relief of policymakers worried about the possibility of being replaced by algorithms!

It is worth saying that this paper has not considered some data sources that may support more 'futuristic' visions of innovation policy, such as 'living labs' that collect large volumes of data through smartphones and sensors, or Smart city data (Pentland, 2015). Perhaps these will start playing a role in innovation policy once the agenda of data analytics experimentation based on the data sources we have considered in this paper have started to show their value, but this is work in progress.

The speed with which all this happens will depend, to a great extent, on the ability of practitioners in the field to dispel uncertainty about the potential applications of data analytics in innovation policy, and on the development of complementary skills, institutions and processes. As the innovation studies literature tells us, this will require a judicious mix of experimentation and coordination, and much knowledge sharing. Perhaps innovation policymakers who are familiar with the dynamics of new technologies and sectors will be able to apply the new ideas we have described in this paper more rapidly than those in other domains.

References

Abreu, M., Grinevich, V., Kitson, M., Savona, M., 2010. Policies to enhance the "hidden innovation" in services: evidence and lessons from the UK. Serv. Ind. J. 30, 99–118. doi:10.1080/02642060802236160

Aghion, P., Blundell, R., Griffith, R., Howitt, P., Prantl, S., 2009a. The Effects of Entry on Incumbent Innovation and Productivity. Rev. Econ. Stat. 91, 20–32. doi:10.1162/rest.91.1.20

Aghion, P., David, P.A., Foray, D., 2009b. Science, technology and innovation for economic growth: Linking policy research and practice in "STIG Systems." Res. Policy, Special Issue: Emerging Challenges for Science, Technology and Innovation Policy Research: A Reflexive Overview 38, 681–693. doi:10.1016/j.respol.2009.01.016

Aharonson, B.S., Schilling, M.A., 2016. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. Res. Policy 45, 81–96. doi:10.1016/j.respol.2015.08.001

Aldrich, H.E., Fiol, C.M., 1994. Fools Rush in? The Institutional Context of Industry Creation. Acad. Manage. Rev. 19, 645–670. doi:10.5465/AMR.1994.9412190214

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete [WWW Document]. WIRED. URL http://www.wired.com/2008/06/pb-theory/ (accessed 7.22.16).

Bakhshi, H., Bravo-Biosca, A., Mateos-Garcia, J., 2014a. Inside the Datavores: Estimating The Effect Of Data And Online Analytics On Firm Performance. Nesta, London.

Bakhshi, H., Edwards, J., Roper, S., Scully, J., Shaw, D., Morley, L., Rathbone, N., 2013. Creative credits: a randomized controlled industrial policy experiment [WWW Document]. URL https://research.aston.ac.uk/portal/en/researchoutput/creative-credits(1d5bf2c8-a00e-4362-834e-2864cfae0bca).html (accessed 7.28.16).

Bakhshi, H., Mateos-Garcia, J., Whitby, Andrew, 2014b. Model Workers. Nesta.

Bakhshi, H., Schneider, P., Walker, C., 2009. Arts and Humanities Research in the Innovation System: The UK Example. Cult. Sci. J. 2.

Barré, R., 2010. Towards socially robust S&T indicators: indicators as debatable devices, enabling collective learning [WWW Document]. URL http://rev.oxfordjournals.org/content/19/3/227.short (accessed 7.22.16).

Blei, D.M., 2012. Probabilistic Topic Models. Commun ACM 55, 77–84. doi:10.1145/2133806.2133826

Borrás, S., Edquist, C., 2013. The choice of innovation policy instruments. Technol. Forecast. Soc. Change 80, 1513–1522. doi:10.1016/j.techfore.2013.03.002

Breitzman, A., Thomas, P., 2015. The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems. Res. Policy 44, 195–205. doi:10.1016/j.respol.2014.06.006

Brin, S., Page, L., 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. Comput. Netw., The WEB we live in 56, 3825–3833. doi:10.1016/j.comnet.2012.10.007

Brynjolfsson, E., Hitt, L.M., Kim, H.H., 2011. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? (SSRN Scholarly Paper No. ID 1819486). Social Science Research Network, Rochester, NY.

Collaboration, O.S., 2015. Estimating the reproducibility of psychological science. Science 349, aac4716. doi:10.1126/science.aac4716

Conway, D., 2010. The Data Science Venn Diagram [WWW Document]. Drew Conway. URL http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (accessed 5.12.14).

Cooke, P., Gomez Uranga, M., Etxebarria, G., 1997. Regional innovation systems: Institutional and organisational dimensions. Res. Policy 26, 475–491. doi:10.1016/S0048-7333(97)00025-5

Crawford, K., 2013. The Hidden Biases in Big Data [WWW Document]. Harv. Bus. Rev. URL https://hbr.org/2013/04/the-hidden-biases-in-big-data (accessed 7.22.16).

Crawford, K., others, 2011. Six provocations for big data.

Cunningham, P., Ramlogan, R., 2012. The Effects of Innovation Network Policies (No. 12/04), Nesta Working Paper. Nesta, London.

Davidson, S., Potts, J., 2016. A New Institutional Approach to Innovation Policy. Aust. Econ. Rev. 49, 200–207. doi:10.1111/1467-8462.12153

Department for Business, Innovation and Skills, 2014. Estimating the effect of UK direct public support for innovation [WWW Document]. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/369650/bis-14-1168-estimating-the-effect-of-uk-direct-public-support-for-innovation-bis-analysis-paper-number-04.pdf (accessed 7.21.16).

Diebold, F.X., 2012. On the Origin(s) and Development of the Term “Big Data” (SSRN Scholarly Paper No. ID 2152421). Social Science Research Network, Rochester, NY.

Dodgson, M., Gann, D.M., Phillips, N., 2013. The Oxford Handbook of Innovation Management. OUP Oxford.

Edler, J., Cunningham, P., Gök, A., Shapira, U. of M. and P., 2016. Handbook of Innovation Policy Impact. Edward Elgar Publishing.

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

Encaoua, D., Hall, B.H., Laisney, F., Mairesse, J., 2013. The Economics and Econometrics of Innovation. Springer Science & Business Media.

Etzkowitz, H., Leydesdorff, L., 2000. The dynamics of innovation: from National Systems and “Mode 2” to a Triple Helix of university–industry–government relations. Res. Policy 29, 109–123. doi:10.1016/S0048-7333(99)00055-4

Fagerberg, J., Mowery, D.C., Nelson, R.R., 2006. The Oxford handbook of innovation. Oxford Handbooks Online.

Flanagan, K., Uyarra, E., 2016. Four dangers in innovation policy studies – and how to avoid them. Ind. Innov. 23, 177–188. doi:10.1080/13662716.2016.1146126

Galindo-Rueda, F., Millot, V., 2015. Measuring Design and its Role in Innovation (OECD Science, Technology and Industry Working Papers). Organisation for Economic Co-operation and Development, Paris.

Giuliani, E., Pietrobelli, C., 2011. Social Network Analysis Methodologies for the Evaluation of Cluster Development Programs. Inter-American Development Bank.

Gök, A., Waterworth, A., Shapira, P., 2014. Use of web mining in studying innovation. Scientometrics 102, 653–671. doi:10.1007/s11192-014-1434-0

Gustafsson, R., Autio, E., 2011. A failure trichotomy in knowledge exploration and exploitation. Res. Policy 40, 819–831.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D.T., Ward, M.D., 2015. Data Science in Statistics Curricula: Preparing Students to "Think with Data." Am. Stat. 69, 343–353. doi:10.1080/00031305.2015.1077729

Hekkert, M.P., Suurs, R.A.A., Negro, S.O., Kuhlmann, S., Smits, R.E.H.M., 2007. Functions of innovation systems: A new approach for analysing technological change. Technol. Forecast. Soc. Change 74, 413–432. doi:10.1016/j.techfore.2006.03.002

Hidalgo, C.A., n.d. What's Wrong with Open-Data Sites--and How We Can Fix Them [WWW Document]. Sci. Am. Blog Netw. URL http://blogs.scientificamerican.com/guest-blog/what-s-wrong-with-open-data-sites-and-how-we-can-fix-them/ (accessed 7.21.16).

Hood, W., Wilson, 2001. The Literature of Bibliometrics, Scientometrics, and Informetrics. Scientometrics 52, 291–314. doi:10.1023/A:1017919924342

Horton, J.J., Tambe, P., 2015. Labor Economists Get Their Microscope: Big Data and Labor Market Analysis. Big Data 3, 130–137. doi:10.1089/big.2015.0017

Hughes, A., Kitson, M., 2012. Pathways to impact and the strategic role of universities: new evidence on the breadth and depth of university knowledge exchange in the UK and the factors constraining its development. Camb. J. Econ. 36, 723–750. doi:10.1093/cje/bes017

Klein Woolthuis, R., Lankhuizen, M., Gilsing, V., 2005. A system failure framework for innovation policy design. Technovation 25, 609–619. doi:10.1016/j.technovation.2003.11.002

Lane, J., Owen-Smith, J., Rosen, R., Weinberg, B., 2014. New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value (No. 8556), IZA Working Paper.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science 343, 1203–1205. doi:10.1126/science.1248506

Malerba, F., 2002. Sectoral systems of innovation and production. Res. Policy, Innovation Systems 31, 247–264. doi:10.1016/S0048-7333(01)00139-1

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity.

Manyika, J., Chui, M., Farrell, D., Kuiken, S.V., Groves, P., Doshi, E.A., n.d. Open data: Unlocking innovation and performance with liquid information | McKinsey & Company [WWW Document]. URL http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information (accessed 7.21.16).

Mateos-Garcia, J., Bakhshi, H., Windsor, G., 2014. Skills of the Datavores: Talent and the data revolution | Nesta [WWW Document]. URL http://www.nesta.org.uk/publications/skills-datavores-talent-and-data-revolution (accessed 7.22.16).

Mathew, M., 2015. Introduction to the Special Section on Patent Analytics. Int. J. Innov. Technol. Manag. 12, 1502001. doi:10.1142/S0219877015020010

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that Will Transform how We Live, Work, and Think. Houghton Mifflin Harcourt.

Mazzucato, M., 2015. The Entrepreneurial State: Debunking Public vs. Private Sector Myths. Anthem Press.

Medina, E., 2011. Cybernetic Revolutionaries: Technology and Politics in Allende's Chile. MIT Press.

Mulas, V., Minges, M., Applebaum, H.R., 2015. Boosting tech innovation ecosystems in cities : a framework for growth and sustainability of urban tech innovation ecosystems (No. 100899). The World Bank.

Mulgan, G., Tucker, S., Ali, R., Sanders, B., 2007. Social Innovation: What it is, why it matters and how it can be accelerated [WWW Document]. URL http://eureka.sbs.ox.ac.uk/761/ (accessed 7.28.16).

Nathan, M., Rosso, A., 2015. Mapping digital businesses with big data: Some early findings from the UK. Res. Policy, The New Data Frontier 44, 1714–1733. doi:10.1016/j.respol.2015.01.008

Nelson, R.R., 1993. National Innovation Systems: A Comparative Analysis. Oxford University Press.

Nesta, 2016. Using Research Evidence. A Practice Guide. London: Nesta/Alliance for Useful Evidence.

OECD, 2014. Data-driven Innovation for Growth and Well-being. Interim Synthesis Report. OECD, Paris.

O'Reilly, T., 2007. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (SSRN Scholarly Paper No. ID 1008839). Social Science Research Network, Rochester, NY.

Pasquale, F., 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Patil, D.J., 2011. Building data science teams - O'Reilly Radar [WWW Document]. URL http://radar.oreilly.com/2011/09/building-data-science-teams.html (accessed 5.6.14).

Peng, R.D., 2011. Reproducible Research in Computational Science. Science 334, 1226–1227. doi:10.1126/science.1213847

Pentland, A., 2015. Social Physics: How Social Networks Can Make Us Smarter. Penguin Publishing Group.

Piwowar, H., 2013. Altmetrics: Value all research products. Nature 493, 159–159. doi:10.1038/493159a

Potts, J., Roe, G., Henderson, B., 2016. Detecting New Industry Emergence Using Government Data: A New Analytic Approach to Regional Innovation Policy (SSRN Scholarly Paper No. ID 2763978). Social Science Research Network, Rochester, NY.

Priem, J., Taraborelli, D., Groth, P., Neylon, C., 2010. Altmetrics: a manifesto [WWW Document]. URL http://altmetrics.org/manifesto/ (accessed 7.22.16).

Rudin, C., Wagstaff, K.L., 2013. Machine learning for science and society. Mach. Learn. 95, 1–9. doi:10.1007/s10994-013-5425-9

Smith, S., Ward, V., House, A., 2011. "Impact" in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. Res. Policy 40, 1369–1379. doi:10.1016/j.respol.2011.05.026

Stirling, A., 2014. Towards Innovation Democracy? Participation, Responsibility and Precaution in Innovation Governance. (SSRN Scholarly Paper No. ID 2743136). Social Science Research Network, Rochester, NY.

Tambe, P., 2013. Big Data Investment, Skills, and Firm Value. Ski. Firm Value May 8 2013.

Tech City UK/Nesta. 2016. Tech Nation 2016: Transforming UK Industries. London: Tech City UK/Nesta

Tether, B.S., 2002. Who co-operates for innovation, and why: An empirical analysis. Res. Policy 31, 947–967. doi:10.1016/S0048-7333(01)00172-X

Tether, B.S., Tajar, A., 2008. Beyond industry–university links: Sourcing knowledge for innovation from consultants, private research organisations and the public science-base. Res. Policy 37, 1079–1095. doi:10.1016/j.respol.2008.04.003

Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C.R., 2013. Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLOS ONE 8, e64841. doi:10.1371/journal.pone.0064841

Ubaldi, B., 2013. Open Government Data (OECD Working Papers on Public Governance). Organisation for Economic Co-operation and Development, Paris.

Von Hippel, E., Jong, J., Flowers, S., 2012. Comparing Business and Household Sector Innovation in Consumer Products: Findings from a Representative Study in the United Kingdom. Manag. Sci. 58, 1669–1681. doi:10.1287/mnsc.1110.1508

World Bank (2014), Open data for economic growth, Transport & ICT Global Practice, http://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf