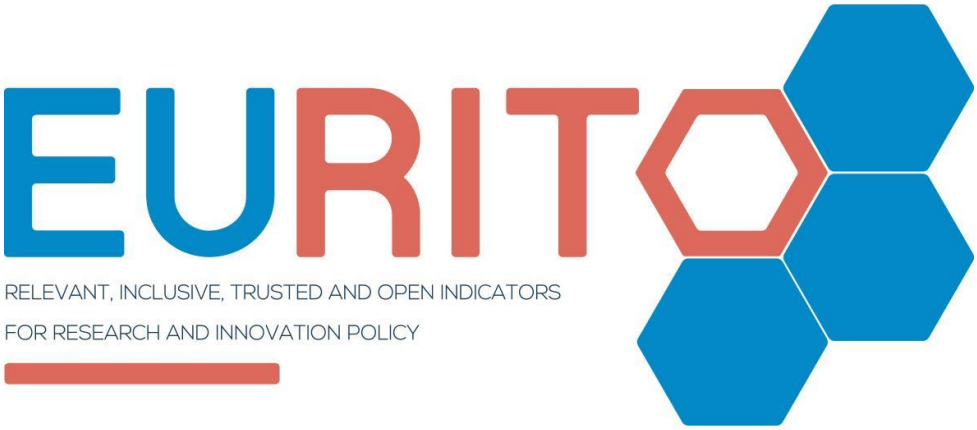


D2.2 Pilots Research Results

Deliverable No.		D2.2	
Workpackage No.	2	**Workpackage Title**	Exploration
Lead beneficiary		Nesta	
Dissemination level		Public	
Type		ORDP: Open Research Data Pilot	
Due Date		M14 (28 February 2019)	
Version No.		0.2	
Submission Date		28 February 2018	
File Name		D2.2 Pilots Research Results 0.1	
Project Duration		36 Months	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 770420.

Version Control

Version	Date	Author	Notes
0.1	2 April 2018	Nesta	Template Creation
0.2	20 February 2018	All Partners	Collation of Reports
0.3	26 February	All Partners	Final Internal Review
0.4			

Reviewers List

Version	Date	Reviewers	Notes
0.2.1			
0.2.2			

Disclaimer

This document has been produced with the assistance of the European Union. The contents of this publication are the sole responsibility of the author and can in no way be taken to reflect the views of the European Union.

Executive Summary

The EURITO project will develop new Relevant Inclusive Trusted Open (RITO) indicators for Research and Innovation (R&I) policy in the EU. In the Exploratory Phase of the project, we have taken eight policy-relevant questions identified during the scoping phase of the project and carried out a systematic exploration of available data sources and methods that can be used to develop indicators addressing them. Our goal is to identify high potential, technically feasible ideas that we can scale up, validate and visualise in the next stage of the project, and to create a data, metadata and software code infrastructure allowing us to do so. In this report, we present the rationale, methodology and emerging findings for each of these exploratory pilots. They are thus:

Pilot 1: Emerging Tech Ecosystems (Artificial Intelligence): This pilot combines data from multiple open and web sources which we analyse using natural language processing to map technology ecosystems in the EU, including their evolution, national and regional geography and industrial relevance. We generate prototype indicators for Artificial Intelligence, an emerging technology of substantial interest to policymakers where the evidence base is currently lacking.

Pilot 2: Nowcasting Business Research & Development: We explore new data sources and modelling strategies to nowcast Business Enterprise Research Development (BERD) at the national and firm level drawing on the EU R&D Scoreboard. We show some challenges with using novel data sources with irregular business coverage for this task, and showcase the potential of Bayesian methods and novel variables such as Economic Complexity Indices for nowcasting this policy relevant variable.

Pilot 3: Technological Change Indicators: We calculate indicators of structural technological change based on 'semantic' discontinuities in topic co-occurrence in a variety of R&D-related data sources. We show that our approach can be used to analyse the historical evolution of the biofuel field, suggesting that this approach could be used in real time to monitor important breaks in technological trajectories suggesting the emergence of new ideas, applications or actors which might be of interest to innovation policymakers and research funders.

Pilot 4: Standards As Innovation Diffusion Indicators: We use data from the International Standards Organisation (ISO) about the adoption of various technology standards in business in order to monitor technology diffusion in companies, and validate this data with a firm-level dataset scraped from German company websites.

Pilot 5: Evidence Base For Mission-Oriented Research & Innovation: We develop indicators for the analysis of the situation, evolution and composition of 'mission fields' that policymakers seek to boost through proactive policies aimed at steering technological innovation in societally beneficial directions. We pilot these indicators in a database of UK publicly funded research projects, focusing on the UK Government's mission to deploy AI to transform the prevention, diagnosis and treatment of chronic diseases.

Pilot 6: Advanced Research & Innovation Funding Analytics: We explore opportunities to apply network science to develop indicators of research impact that are more context-aware and able to capture the way in which innovative research transforms the structure of knowledge networks through new recombinations of topics. We build prototypes of these

indicators using data about EU funded research projects and outputs from CORDIS and the OpenAIRE repository.

Pilot 7: Inclusive Innovation: We use a machine learning algorithm to predict gender and ethnicity of technology company personnel in the CrunchBase company directory. We use this information to produce indicators of socio-demographic diversity in different countries, technology sectors and cities, also considering intersectional metrics capturing specific combinations of socio-demographic attributes.

Pilot 8: Linkages and Knowledge Exchange Indicators (Healthtech): We map the network structure of research projects and outputs in the funding data of the Novo Nordisk foundation and the CORDIS and OpenAIRE datasets. We generate indicators helping us understand the link between research funding and collaboration and knowledge exchange in the EU.

The range of research questions, data sources and methodologies featured in our pilot portfolio demonstrate the breadth of opportunities to use new data sources and methods to develop RITO indicators for innovation policy in the EU. We will continue exploring these opportunities at a larger scale for a subset of the pilots in the next stage of the project.

Pilot 1: Emerging Technology Ecosystems

Abstract

Emerging technologies with strong market potential and prospects for improving productivity are of great interest for Research and Innovation (R&I) policymakers, but it is difficult to monitor their development and diffusion using pre-structured categories and aggregated data sources. In this pilot, we explore an alternative approach to generate indicators about emerging technology R&D and its technological innovation system using a collection of unstructured data sources that we query with Clio, an information retrieval tool. We do this focusing on Artificial Intelligence (AI), a potentially transformational technology where countries across the world are making substantial investments but where, until recently, the evidence base has been weak. We set out work in the context of recent efforts to map AI R&D and present illustrative indicators about levels of AI research, business, open source software development and informal networking activity in the EU compared to other territories, the geographical distribution of AI activity in EU regions, the composition of different EU AI systems, and the industrial orientation of AI activity in different data sources and countries. We conclude by triangulating our indicators with official sources, reviewing the strengths and weaknesses of the approach, and considering options for scaling-up the analysis.

1 Introduction

1.1 Background/context

Emerging technologies with rapid growth potential and widespread applicability are increasingly garnering the attention of R&I policymakers (Rotolo, Hicks, & Martin, 2015). Some examples include Artificial Intelligence, Immersive technologies such as Virtual or Augmented Reality, Blockchain and distributed ledger technologies, and robots and drones. Policymaker interest in these technologies is motivated by the desire to gain a foothold in sectors with fast growth potential and first-mover advantages, to encourage the adoption of productivity enhancing technologies in existing industries, and to steer their development in societally desirable directions (Aghion, David, & Foray, 2009; Stilgoe, Owen, & Macnaghten, 2013). Emerging, general-purpose technologies could also play an important role in the delivery of mission-driven innovation policies (Mazzucato, 2018).

The design and development of programmes and instruments to achieve these goals are however hampered by lack of Relevant, Inclusive, Trusted, Timely and Open (RITO) indicators about the emergence and diffusion of new technologies, and about the availability of complementary resources such as skills, finance, informal networks or infrastructure which are required for their successful deployment (Börner, Rouse, Trunfio, & Stanley, 2018; Börner, Scrivner, et al., 2018).

The reason for these gaps in the evidence base is that novel, emerging technologies are almost by definition missing from pre-established taxonomies used for industrial, scientific and technological analysis that tend to provide a lagging perspective on the economy (Bakhshi & Mateos-Garcia, 2016). The widespread applicability of these technologies also means that many of them are components or tools for the delivery of new products, services and research activities, yet this diffusion is hard to capture in mutually exclusive, completely exhaustive industrial and scientific taxonomies that shoehorn organisations in a single sector regardless of the technologies they use (Hicks, 2011; Nathan & Rosso, 2015).

Novel, unstructured data sources such as repositories of scientific research and research funding, the text in patent abstracts, business descriptions in company directories and websites, and information from online platforms used to coordinate new technology development and distribute its outputs could help fill these gaps in the evidence base (Bakhshi & Mateos-Garcia, 2016; Börner, Rouse, et al., 2018). Many of these sources contain rich textual descriptions of the content of the R&D and innovative activities being undertaken by actors which can be mined using a variety of Natural Language Processing (NLP) approaches in order to identify entities (papers, projects, patents, companies) related to an emerging technology that can be used to generate novel indicators to inform policy and practice (Hain & Jurowetzki, n.d.; Jurowetzki & Hain, 2014). Increasingly, there has been an expansion of research in this area.

However, these new approaches also present challenges. Novel data sources suffer from representativity biases, dynamic inconsistencies (as the levels of activity inside them could reflect changes in their popularity and design) and lack of robustness if the results are sensitive to changes to the parameters used by the analyst (Salganik, 2017). Building trust around these methods requires using, as much as possible, open data sources, triangulation between data sources and reproducible analyses so that other researchers can review the work and identify potential weaknesses (Peng, 2011). Unfortunately, much work in this area lacks the openness required to undertake this methodological due diligence, making it difficult to build trust on new methods and identify standard practices for the identification and monitoring of emerging technologies. Additionally open data sources could lack consistency to perform a continuous monitoring of the data (Lazer, Kennedy, King, & Vespignani, 2014).

1.1.1 Opportunity

The pilot we describe in this paper seeks to address some of the challenges above by using novel, unstructured data sources that characterise different activities and components in the technological system for an important emerging technology, AI. As part of the pilot, we are developing an open and reproducible infrastructure for data collection, processing and analysis that other researchers can use to validate our work in ways that build trust around it. Ultimately, we seek to create an interactive system enabling policymakers and practitioners to flexibly identify organisations and projects related to an emerging technology, and generate RITO indicators about them that can inform policy.

1.1.2 Application domain

We test our approach in the application domain of Artificial Intelligence (AI). AI is a collection of technologies that use data and machine learning methods to generate predictions to automate and/or inform economic activities (A. K. Agrawal, Gans, & Goldfarb, 2018b, 2018a; Mateos-Garcia, 2018).

In recent years, increasing availability of data, improvements in data storage and computing, and innovations in algorithms for deep and reinforcement learning have significantly improved the performance of AI systems, resulting in new applications in computer vision and speech synthesis, information retrieval, unmanned vehicles and robotics (J. Klinger, Mateos-Garcia, & Stathoulopoulos, 2018).[1] A growing number of economists argue that AI is the latest example of a "General Purpose

[1] Deep learning algorithms are based on networks where data are processed by subsequent layers where artificial neurons create increasingly abstract representations of the data which correspond to real-world characteristics of the objects being represented (Goodfellow, Bengio, & Courville, 2016). Reinforcement learning involves using reward functions that 'teach' algorithms how to achieve goals Through trial and error in interactive and synthetic environments (Arulkumaran, Deisenroth, Brundage, & Bharath, 2017).

Technology" (GPT) that will transform economies, societies and international relations (A. Agrawal, McHale, & Oettl, 2018; Cockburn, Henderson, & Stern, 2018; Goldfarb & Trefler, 2018; Levy, 2018). As it has been the case with other GPTs such as steam or electricity, the successful deployment of AI will require investments in technology development and diffusion to be accompanied with an increase in the supply of workers with relevant technical skills, and changes in production processes and business models, and regulatory regimes (Trajtenberg, 2018). The systemic nature of AI, like other emerging technologies, make it important to consider these complementary inputs, activities and framework conditions in order to understand drivers and barriers to its successful deployment.

Policy and public interest in AI echoes its expected, substantial, impacts, and governments across the world have put in place national strategies to support homegrown AI industries, encourage the deployment of AI to enhance productivity in the private and public sector, and steer its development and application in ways that are safe and respect privacy and human rights (Goldfarb & Trefler, 2018). The European Union has been an early mover in this space, with its European Initiative on AI for Europe committing to invest €1.5 Bn on AI between 2018-2020, including through research funding in the Framework Programmes and investments on Digital Innovation Hubs to encourage the diffusion of AI into industry (European Commission, 2018). Meanwhile, the General Data Protection Regulation (GDPR) and a recently formed High Level Expert Group for AI are putting the EU at the forefront of ethical AI development which is aligned with European values. This concerted effort is timely, given the prevailing (if recently challenged) perception that the EU is lagging behind other territories such as the US and China in the development of strong AI industries .

Until recently, the evidence base for these AI policies was piecemeal and fragmented, drawing on case studies and information from market research firms and commercial consultants. This situation has only recently started to be remedied with the publication of several AI mapping efforts using more rigorous and sophisticated methodologies for the identification and analysis of AI R&D trajectories. We overview them in the methodology section of this paper, comparing them with the approach that we follow in our own work.

1.1.3 Flexibility of the application domain

All the data sources and methods we use in the pilot are domain-agnostic and could in principle be used to analyse other emerging technologies, with the proviso that some of our data sources, such as CrunchBase (a start-up directory) or GitHub (an open source software repository) will have better coverage of digital emerging technologies than, for, say, emerging technologies in manufacturing or pharmaceuticals. We briefly indicate potential biases in the coverage of different data sources when we present our methodology.

1.1.4 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

As we mentioned above, emerging technology indicators based on new sources have only recently started to gain currency in policy discussions and are far from being used regularly, beyond providing empirical context for R&I strategies. None of the indicators in the European Innovation Scoreboard, the central framework for R&I monitoring and benchmarking in the EU, consider emerging technologies, focusing instead on aggregate research and technology development through publications and patenting, and activities in sectors which are pre-defined as 'innovative', 'knowledge intensive' or 'high-tech'.

In this pilot, we seek to develop indicators that are relatively easy to incorporate within this traditional R&I measurement framework using methodologies that build trust around their use, hopefully contributing to their adoption into the 'core' of R&I policy.

1.1.5 Stakeholder engagement summary

The approach and results presented in this pilot have received feedback from multiple research and policy audiences. Going beyond the EURITO Knowledge Stakeholder Workshop, where we received detailed feedback from policy experts in DG RTD, the Joint Research Centre and the OECD, we have also presented our AI innovation mapping work to Innovate UK, the UK's technology agency, and audiences of leading Science, Technology and Innovation scholars at SPRU in the University of Sussex, and the Santa Fe Institute. The technical and policy input that we have received is reflected throughout the report and specially in the conclusions, when we consider strengths and weaknesses of our approach and next steps for the pilot.

1.2 Relevance to RITO criteria

1.2.1 Relevant

The work being undertaken in this pilot is highly policy-relevant, given the chasm between R&I policy interests in emerging technologies and the lack of robust evidence to inform and baseline the policies being put in place. The application domain for the pilot, AI, is particularly topical, with dozens of AI national strategies committing billions of Euros to AI development in recent years based on a weak, fragmented or even anecdotal evidence base.

1.2.2 Inclusive

The indicators that we develop here seek to achieve inclusivity in two ways.

First, we consider geographical inclusion by looking at the spatial distribution of AI activity inside countries. This way, we aim to evidence the concern that the development and diffusion of AI technologies could be worsening the economic divide between 'creative/tech' cities developing AI technologies that automate the workforces of 'left-behind' places, thus increasing inequality, societal divides and political volatility (Korinek & Stiglitz, 2017; Levy, 2018; Trajtenberg, 2018).

Second, we consider sectoral inclusion by exploring the extent to which AI R&D is 'semantically close' (i.e. applicable) to different sectors. This way, we want to explore the hypothesis that an inordinate amount of investment and research effort is being devoted to the development of AI applications in sectors such as digital technology and media at the expense of other domains such as health, education or public service delivery where these technologies could also create significant impacts (Mateos-Garcia, 2018; Mulgan, 2017).

1.2.3 Timely

Almost without exception, the data sources that we work with are available close to real time and could therefore be used to generate highly timely indicators of emerging technology activity and the situation in emerging technological innovation systems.

1.2.4 Trusted

We are making (almost) all the data and code we use in the project open and reproducible so that our analysis can be validated by others. All our code is available in this GitHub repository: https://github.com/nestauk/eurito_pilot_1_emergent. We are also obtaining feedback from domain experts about our emerging findings, and triangulating our findings with related data from official sources such as Eurostat.

1.2.5 Open

All the datasets we are currently using in the project are open with the exception of CrunchBase, a startup directory. This means that we will be able to make the majority of the data we use available to enable extensive validation and peer review of our work, and follow-on analyses.

1.3 Research/policy questions

1. What are the levels of AI R&D in the EU compared to 'competitor' territories, and how has it evolved over time?
    a. Are there any qualitative differences between the types of AI R&D undertaken in the EU and 'competitor' territories?
2. What are the levels of AI R&D in EU member states and how has it evolved over time?
    a. What is the structure of different AI technology ecosystems inside the EU?
3. What is the regional distribution of AI R&D activity inside the EU and how does it compare to R&D activity in general, and to the distribution of the population?
4. What is the sectoral focus of AI R&D and how does it diverge between EU member states?

2 Methodology

In this section we overview the data sources we have considered / used in the pilot, and how we collected and processed this data, paying special attention to the approach we have used to identify AI R&D activities. We indicate potential problems and challenges as we go, also comparing our approach to recent AI mapping work by other research teams (see table 1 for an outline of those efforts).

Table 1: Recent AI maps

Authors	Data sources	AI detection method / level of analysis	Findings
(Cockburn et al., 2018)	Publications (Scopus) and USPTO Patents up to 2014	1. Keyword search in abstracts. Distinguishes 'deep learning' from 'symbolic AI' and robotics. Aggregate analysis.	• 98K AI publications (58K 'learning') and 13K AI patents (3.8K 'learning'). • Rapid growth in activity, ML spreading between disciplines. • USA was lagging behind but is now catching up.
(Mann &	Patents and	1. Label automation patents and	• 2.2m out of 5 million

Püttmann, 2017)	industries	train a supervised machine learning model. 2. Identify industries exposed to automation through patent code - industry code lookup.	'automation' patents.
(Elsevier, 2018; Siebert, Kohler, Scerri, & Tsatsaronis, 2018)	Publications, patents, university sylaby and news	1. AI vocabulary extracted from selected sources and validated by domain experts 2. Scopus queried with AI vocabulary. 3. Subset labelled by experts and used to train a supervised machine learning model to predict likelihood that paper is AI. 4. Topic co-occurrence analysis to identify AI sub-fields. 5. Use of OECD FORD categories to identify domain relevance of papers (method unclear). Analysis by territory and country	● 600K 'AI' documents. ● Rapid growth ● EU has most activity ● Differences in focus between territories (eg. China more focused in computer vision, EU more focused in robotics).
(AI Index, 2017, n.d.)	Papers, course enrollment, jobs, downloads, performance improvements etc.	1. Depends on data source Analysis generally aggregate with some metrics reported by country.	● Paper / patent analysis draws on Siebert et al above. ● Fast growth in activity
(J. Klinger et al., 2018)	arXiv and CrunchBase	1. Topic modelling of paper abstracts and identification of two Deep Learning related topics. 2. Identify industry relevance of arXiv papers with a model trained on CrunchBase data. Analysis by country and NUTS-2 region	● 15K deep learning papers. ● Rapid growth in activity ● EU has lost competitiveness versus China and US.
(Joint Research Centre,	Companies, conference submissions,	1. Keyword search in descriptions, abstracts etc. 2. Use of topic modelling to	● 35,000 AI 'players' globally (25% in EU).

2018)	patents, startups, investment, university courses	identify sub-fields of interest. Analysis by territory and country.	

2.1 Data sources

Our pilot is organised around the concept of an 'technological innovation system' . This notion, which is strongly connected to influential concepts such as National Systems of Innovation, Sectoral Systems of Innovation, innovation collectives, and Entrepreneurial Ecosystems, is defined as the "*the set of actors and rules that influence the speed and direction of technological change in a specific technological area*", thus contributing to its development and diffusion (Bergek, Jacobsson, Carlsson, Lindmark, & Rickne, 2008). It includes:

a) Researchers in academia and the public and private sector that conduct research activities related to an emerging technology, resulting in publications, patents and software tools
b) Businesses that adopt an emerging technology to develop new products and services and improve their processes, thus increasing their productivity, and the trade bodies and business networks
c) Investors and funders that provide finance for these R&D&I activities
d) Workers with the skills and competences to develop and use an emerging technology, as well as groups of workers such as trade unions and professional associations
e) Educational institutions and communities that offer skills on how to develop and use the emerging technology
f) Government actors that procure an emerging technologies, set standards and regulate emerging technology development and application through a variety of channels
g) Consumers who demand (or reject) products and services based on an emerging technologies, and seek to steer their development, generally through the representation of civic society groups

This list is extensive. Collecting data on all these actors and activities goes beyond the scope of the pilot. Table 1 outlines some of the data sources we have considered / used in our analysis until now.

Table 2: Data collection summary and observations

Source	Status	Rationale and observations
Research (arXiv)	Collected and analysed all data in computer science categories from API	A Science, Technology, Engineering and Maths (STEM) pre-prints database widely used by AI engineers. Poor coverage of Arts & Humanities and Social Sciences research.

Research (Microsoft Academic Graph)	To be collected from API	A repository of 80 million papers, books and conference proceeding available through an API, similar in coverage to Scopus / Web of Science.
Technology development (PATSTAT)	Licensed (not yet analysed)	A patent database with 70 million patents from developed and developing country intellectual property offices.
Research funding (H2020)	Collected and analysed all data through CORDIS data download.	A repository with information about EU R&D funding through the H2020 programme running between 2014 and 2020. Mostly covers EU member states, does not include funding from national research funding councils, national ministries of science etc.
Software development (GitHub)	Collected and analysed all projects with location data from data dump stored in Google Big Query.	A collaborative software development and code sharing platform hosting 100 million projects and 30 million unique registered users. The data only covers public projects available with an open source license. It therefore excludes proprietary software.
Business activity (CrunchBase)	Licensed and analysed	Global business directory with information about startups and technology companies hosted and curated by tech news website TechCrunch. Its coverage is better for Western countries, technology sectors and younger companies (Dalle, den Besten, & Menon, 2017).
Skills supply (University websites)	Piloted but not analysed	We previously developed a system that uses network analysis to identify department pages in university websites, and topic modelling to detect clusters of words related to the skills supplied by the department (Joel Klinger et al., 2018). This will not capture other forms of education and training and professional development.
Skills demand (online job adverts through Indeed or Monster)	To be collected	In principle, these data could be used to analyse demand for personnel with AI related skills in different industries and locations. Coverage likely to be better for technical professionals and occupations (which is the relevant population for this pilot).

Skills sharing (informal networking via Meetups)	Collected and analysed all data for EU-based technology meetups and members from Meetup API.	Event platform hosting the activities of 330,000 groups with 39 million members. Coverage likely to be better for digital industries and professions. Meetup activity likely to capture business networking as well as skills and knowledge sharing.

In order to incorporate these data into our analysis we need textual descriptions of the topic for a project, and geographical information to map activity at the national and regional level. In both cases, we find data-specific issues.

Regarding text, there are important differences between the length of descriptions across data sources. For example, arXiv papers have a median description length of 136 words while CrunchBase and GitHub have median description lengths of 14 and 11 words respectively. This could create problems for our approach to identify AI activities based on text descriptions, and the model we use to establish the relevance of activities for different industries. One avenue to address this as/if we scale up the pilot would be to collect additional information about GitHub projects and CrunchBase companies from web sources.

On the topic of geography, we have disparity in coverage across data sources: some of them such as arXiv or CrunchBase have global coverage while others have EU coverage. In the case of H2020 data, this is by design (the majority of organisations eligible to apply for these funds are based in the EU) while in other cases this is linked to the stage of our data collection. For example, we have so far only collected Meetup data for EU countries, but it would be possible to collect additional data for other territories if required.

It is also worth noting that there is variation in the source of geographical information about organisations. In some cases such as H2020 data, CrunchBase or GitHub, address data are available at source, which we have used to geocode organisations/projects at the NUTS-2 level.[2] In other cases such as with arXiv, the geo-coding has been more complex, and involved matching research organisations in the Microsoft Academic Graph database and the Global Research Identifier (GRID). The process is described in more detail in Klinger et al(J. Klinger et al., 2018)).

2.2 Method

Combined, the data sources above include information about just over 7.2 million entities including research papers and EU-funded projects, companies, open source projects and technology meetups. The analytical challenge is to identify AI-related entities in this large database.

As Table 1 shows, there are a variety of approaches to do this, and each has its pros and cons:[3]

[2] In the case of GitHub, an issue for further discussion is the interpretation of the 'location' of a project given the decentralised nature of the communities that collaborate in many open source projects.

[3] In the discussion that follows, the corpus has already been pre-processed. This involves tokenising, part-of-speech tagging, removal of stop words and rare words, and creation of bigrams and tri-grams

1. **Keyword search**: *Compile a lexicon of keywords related to AI and use them to query the data*. This method is easy to interpret but requires deep domain expertise and updates over time as the vocabulary of a field evolves. This approach also suffers from low precision if terms in the vocabulary are used (perhaps with a different meaning) in another domain in the data, and low recall if the initial vocabulary misses relevant terms. This is the approach used by Cockburn et al (2017), JRC (2018) and (as an initial filter) by SIebert et al (2018).
2. **Supervised machine learning**: *Label a subset of the data as 'AI related' and train a machine learning model that identifies features in a project (eg. words in an abstract) that are highly predictive of the AI label. Use that model to predict labels for the unlabelled dataset*. This approach implicitly generates a lexicon from the labelled data, and (depending on the algorithm used) estimates probabilities that an unlabelled entity belongs to the category of interest. As a downside, it requires an upfront effort on labelling a sufficiently large training set. Biases at this stage will degrade the quality of the results. This approach is used by Siebert et al (2018) in a keyword-filtered dataset, and by Mann and Putmann (2017).
3. **Topic modelling:** *Model the topics in a corpus based on word co-occurrence in documents and identify those that are related to AI.* A topic model identifies clusters of interrelated keywords related to a topic reducing, to some extent, the need to generate a comprehensive lexicon ex ante (Blei, Ng, & Jordan, 2003). Since this analysis is unsupervised, it does not require labelling. Many topic models estimate a probability distribution of topics over documents that can be used to select thresholds above which a document is classified in the category (topic) of interest. As a downside, topic modelling outputs can be hard to interpret, lack robustness and are less adaptive to user needs than the approaches above - if the category of interest is not captured by any of the topics fit by the model, there is no way to identify related entities in the data. This approach is used in Klinger et al (2018).

We have developed an open source tool called Clio that seeks to manage some of the trade offs (Stathoulopoulos, forthcoming). Clio is an information retrieval system that takes a query (eg. a keyword of interest such as 'Artificial Intelligence') and looks for its 'synonyms' - words which are close to the initial seed term in a multidimensional space where all words in the corpus are represented as vectors based on their semantic similarity, calculated by the word2vec algorithm (Mikolov, Yih, & Zweig, 2013). In the example above, these could be terms such as 'Machine Learning', 'Deep Learning' etc. that tend to appear in a similar context to 'Artificial Intelligence'. Having generated an 'expanded keyword' set, we extract the documents with any of those keywords.

One risk here is that some of the synonyms of the seed term will be generic (eg the term 'data', in the example above). In this case, the query will return a large number of irrelevant documents. We try to avoid this problem by removing from the expanded term list keywords with low TF-IDF (Term-Frequency Inverse-Document Frequency) scores (the TF-IDF score normalises the number of times that a term appears in a document by the number of times it appears in a corpus - low TF IDF documents tend to be uninformative because they tend to be evenly distributed across the corpus).

As an example, we have used three seed terms to query Clio: "Artificial Intelligence", "Machine learning" and "Deep Learning". This returns a list of around 260,000 AI-related entities including just under 90,000 arXiv papers, 31,000 CrunchBase companies and 137,000 GitHub projects globally and (only for the EU) 4,135 H2020 projects involving 18,500 organisations, and 2,700 meetup groups.

For frequently occurring pairs and trios of unigrams. We refer to these indistinctly as 'words', 'keywords' and 'terms'.

During analysis and stakeholder engagement, we identified some issues with this approach that we will tackle in the scale-up phase:

- Clio does not generate a relevance score for the results. This makes it difficult to adjust the parameters of a query to test its robustness / manage trade-offs between precision and recall.
- We are currently training the model in a corpus that integrates all data sources. This ignores semantic differences between keywords in sub-corpora (i.e. the fact that there might be differences in the position of our seed words in vector space for different corpora - “AI” could be semantically close to different words in arXiv and in Meetup.).
- SImilarly, we are assuming that the term distribution for the integrated corpus is the same as for its constituent sub-corpora. If this assumption is violated (eg certain words concentrate in specific sub-corpora), this could bias the TF-IDF scores we use for filtering.
- Our analysis is static: we do not consider changes in vocabulary over time.
- Our analysis only considers documents in English. This is currently not a big problem because most of the data sources we are using are in English, but it could become an issue if/when we expand our analysis to include multilingual sources (e.g. university websites, company websites, national sources of research funding).

2.3 Documentation

The code developed and employed in this pilot is available in this GitHub repo: http:////github.com/nestauk/eurito_pilot_1_emergent. This includes Jupyter Notebooks to query Clio and to analyse and visualise the data.

3 Results

This section presents some of the indicators that we have developed as part of the pilot, and their findings. In general, we use simple indicators such as counts of activity, changes in levels of activity and shares of total activity.

We also normalise our counts of activity through Relative Advantage Indices (RAI), which measure the global share of activity of interest (eg. AI papers) in a location compared to its global share in all activities. Formally, the RAI is the same as location quotients used in economic geography, and Revealed Comparative Advantage (RCA) indices used in analyses of trade and the economic complexity literature (Hidalgo & Hausmann, 2009). If the value of RAI_{AI} for a location and activity is above 1, this means that the location is relatively specialised in AI because AI is overrepresented in it compared to the average for all locations. This would suggest that the location has some sort of comparative advantage in AI due to the presence of favourable conditions, access to inputs that are conducive to AI, or of actors with a strong AI capability.

RAI is however not without limitations: since it is a normalised measure, it does not take into account the level of activity in a location: a location with 1000 AI companies in a population of 10,000 will have the same RAI_{AI} as a location with 1 AI company in a population of 1,000 companies. Further, the RAI for smaller locations will tend to be noisier because small, potentially random deviations from average levels of activity will have a strong impact on its RAI. RAIs also present some challenges for longitudinal analysis. For example, if a location experiences a decline in overall activity but its decline in AI is slower, then its RAI_{AI} will increase even if it is decreasing in real terms.

For all these reasons, it is generally a good idea to consider RAI together with absolute measures of activity - we follow this approach when we present our results below.

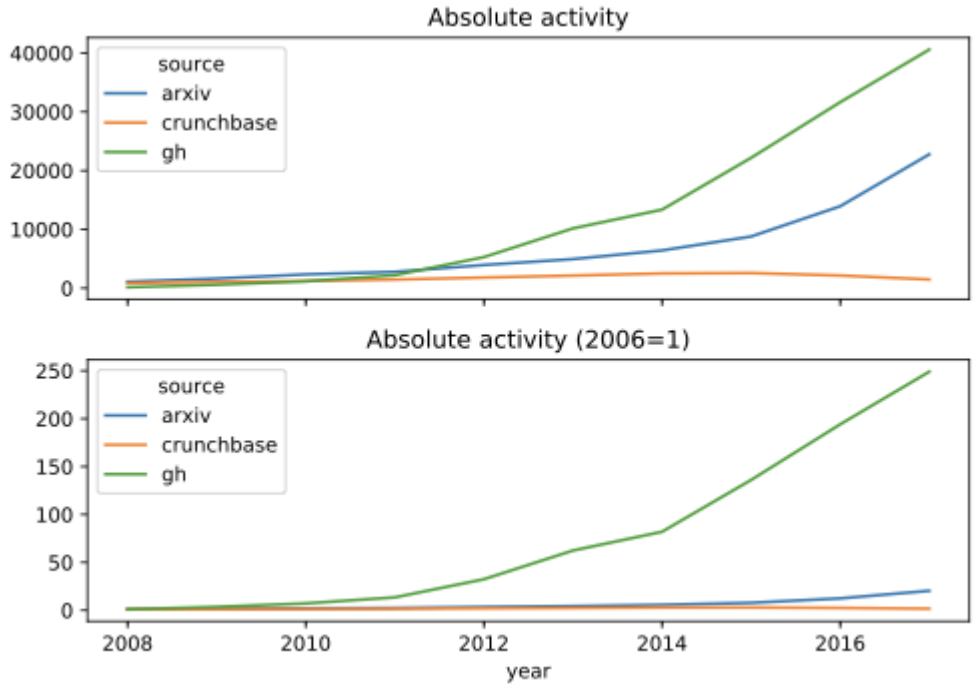
When we look at geographical concentration of AI activity, we use Herfindahl indices which aggregate the square of the share of activity of all locations. A higher Herfindahl score reflects stronger concentration in a small number of locations (if the herfindahl is 1, this means that all activity is located in a single place). One limitation of the Herfindahl index is that it does not consider differences in size between locations: for example, if one region in a country is much larger than the others, we should not be surprised to see more AI activity taking place there than in other regions. We try to account for this by presenting concentration of activity for AI together with other variables (total levels of activity and population).

a. What are the levels of AI R&D in the EU compared to 'competitor' territories, and how has it evolved over time?

We focus first on global levels of AI activity, trends and differences between 'competing territories' using those datasets with global coverage (arXiv, CrunchBase and GitHub).

The two figures below show AI-related activity by data source in absolute terms considering total numbers and using 2006 as an index year. We see a significant increase in activity, particularly in open source activity and arXiv, over the considered period. The 'boom' in AI open source software development eclipses the other two data sources in absolute terms.

Figure 1

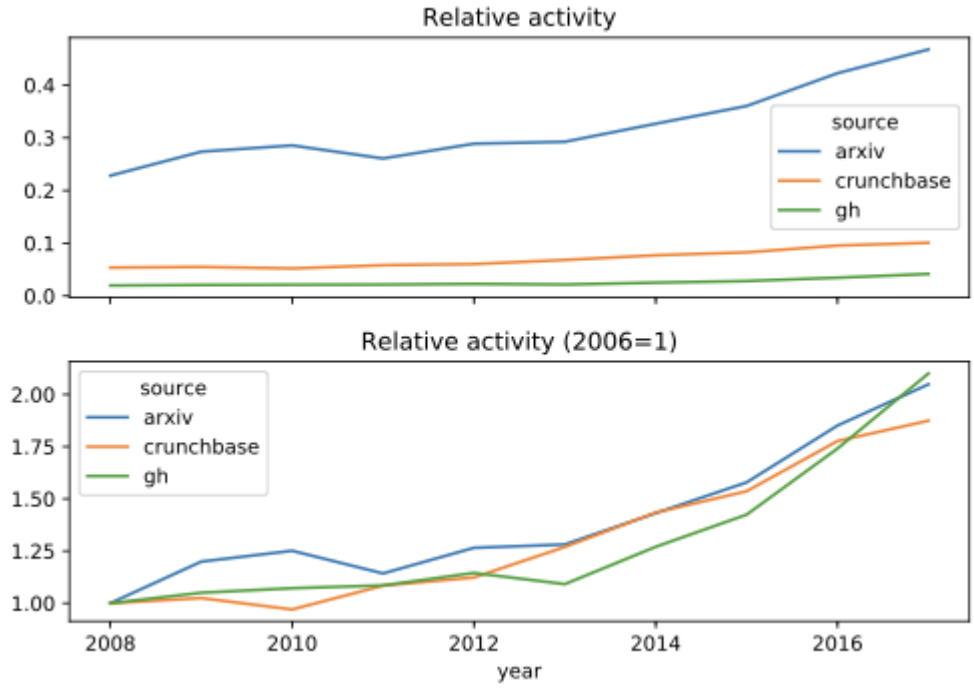


The figure above does not take into account changes in the levels of overall activity in open source software development, publication of pre-prints, and start-ups since 2006 (for example, GitHub was only founded in 2008). To account for this, below we consider levels of AI activity as a share of all activity in a data source.

We see that AI has not only been growing in absolute numbers, but also in relative importance in all the sources that we are considering - this is particularly visible in arXiv, where over 40% of the papers in 2017 were related to AI. The line-chart indexed to 2006 suggests a discontinuity since around 2011, the year of the arrival of Deep Learning, a new machine learning technique that has contributed to important breakthroughs in computer vision, translation and speech synthesis.[4] It is interesting to note that the increase in research activity in arXiv seems to have preceded AI entrepreneurial activity and open source software development, perhaps reflecting a linear model of R&D where scientific breakthroughs eventually lead to the development of new tools and their application in industry.

[4] In 2011, AI systems based on deep learning achieved significant improvements in performance in the ImageNet competition (Krizhevsky, Sutskever, & Hinton, 2012).

Figure 2

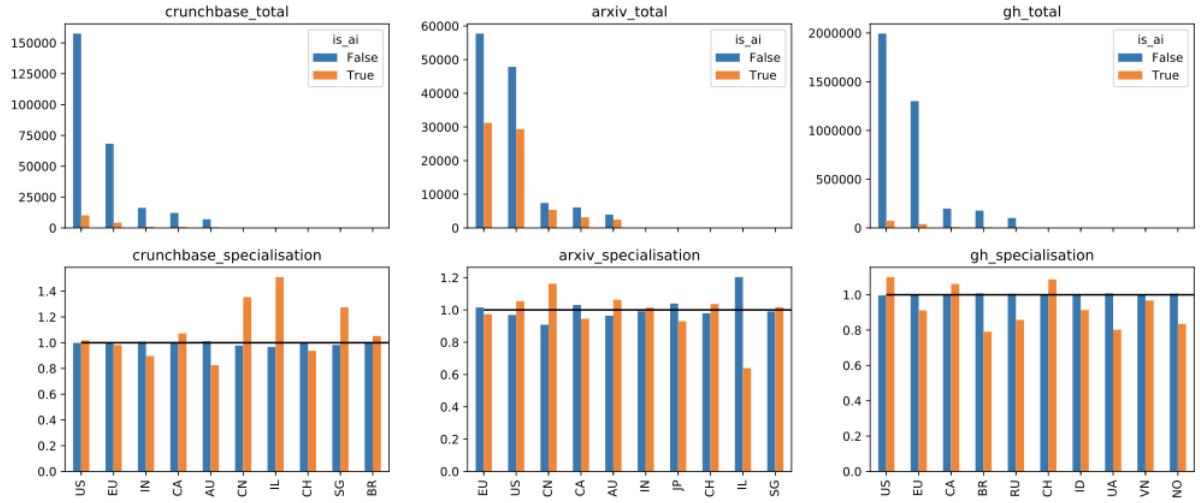


Having considered aggregate levels of activity in AI, we move on to consider international differences. Here, and in common with other recent studies, we aggregate all EU member states into an EU category. In the figure below we present the results in absolute terms (top row) and using the Relative Advantage Index (RAI) we introduced before.

The USA and the EU are dominant in absolute terms in all the datasets that we consider. In line with expectations, China appears underrepresented in the CrunchBase data and more or less absent from the GitHub data (probably due to the presence of competing open coding platforms in China - this is a recurring challenge when using web sources to compare the situation of China with other countries).

The picture changes somewhat when we normalise the data. Then we find that China, Israel, Singapore and Canada are overrepresented amongst the AI startups, China is overrepresented in AI research, and the USA is overrepresented in open source software development. The EU is relatively under-represented in AI research and open source software. Some of the disparities between data sources start to highlight the challenges of using web data for global comparisons: for example, could the low levels of AI research and open source software development activity in Israel reflect different strategies for the dissemination of information amongst the communities of researchers and developers there?

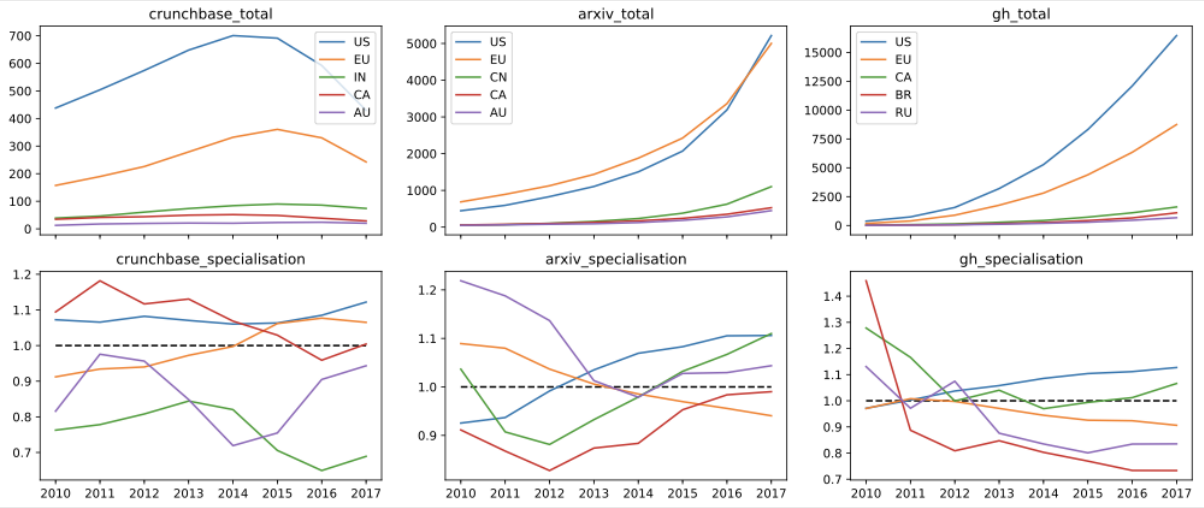
Figure 3



The figures below show the evolution of activity for the top 5 countries by level of activity in each dataset. As below, we show the totals in the first row and the relative advantage index in the second row.

Some notable trends include the apparent loss of initial advantage in AI entrepreneurial activity in Canada, the loss of relative importance of the EU in AI research compared to the US (which initially lagged behind but catches up in line with the Cockburn et al's findings), China and Canada, and the relative collapse of AI open source software development activity in Brazil and Russia. Here it is worth observing that the relative advantage index for the GitHub data might be noisy in the initial periods given low levels of activity in the platform soon after it was founded.

Figure 4



What are the levels of AI R&D in EU member states and how has it evolved over time?

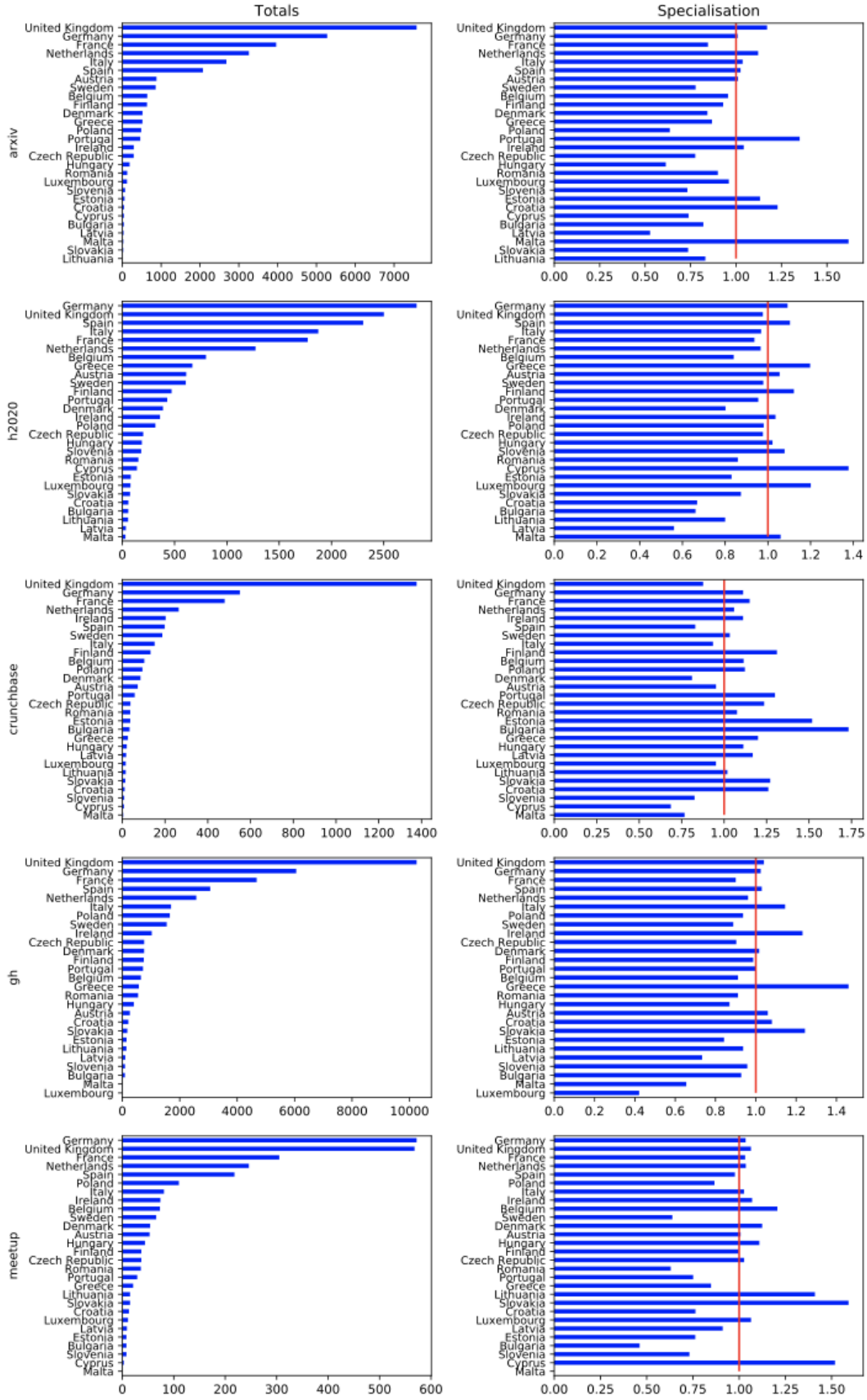
We move on to compare the situation in different EU member states making use of a more extensive collection of datasets, also including H2020 funding data and technology meetups. We continue considering absolute levels of activity and relative advantage.

The figure below presents key findings. It shows that, in absolute terms, the UK is the dominant AI player in the EU in terms of research, entrepreneurship and open source development followed closely by Germany. However, Germany has more organisations receiving funding through H2020 projects, and a slightly larger number of technology meetups related to AI than the UK.

When we look at specialisation levels in the second column, where once again a score above 1 reflects 'relative specialisation' in AI, we begin by noting that the scores for countries with lower levels of activity are likely to be noisier. Having said this, we find some evidence of competitive AI scenes in Estonia and Croatia.

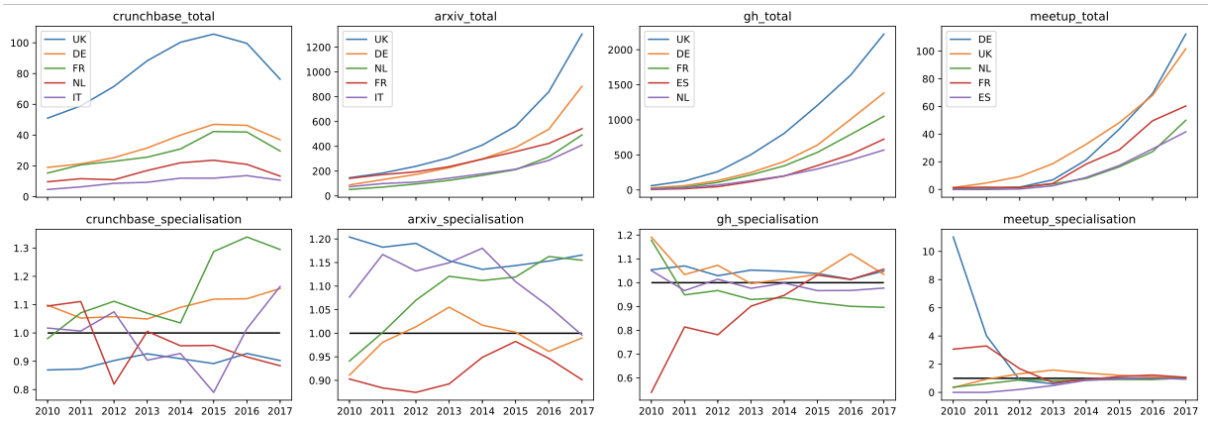
When looking at bigger countries with more activity in absolute terms, United Kingdom and Italy appear to be more competitive in terms of research, Spain and Germany are comparatively better at attracting H2020 funding, Germany, the Netherlands, France and Ireland have a stronger AI startup scene, and Italy, Ireland, Spain and Greece are particularly active in AI open source software development.

Figure 5



We consider once again AI activity trends since 2010 (the beginning of the recent AI 'boom') comparing EU member states. In absolute terms, we see an increase in activity across all dimensions and countries. When we consider specialisation, we detect some interesting patterns, such as an increase in the number of AI-related startups in France, Germany and Italy, a decline in Italy's advantage in research (while Netherlands grows rapidly) and a 'boom' in AI open source activity in Spain. The Meetup figures are highly distorted by noisy relative advantage indices in 2010, very soon after the establishment of the platform.

Figure 6

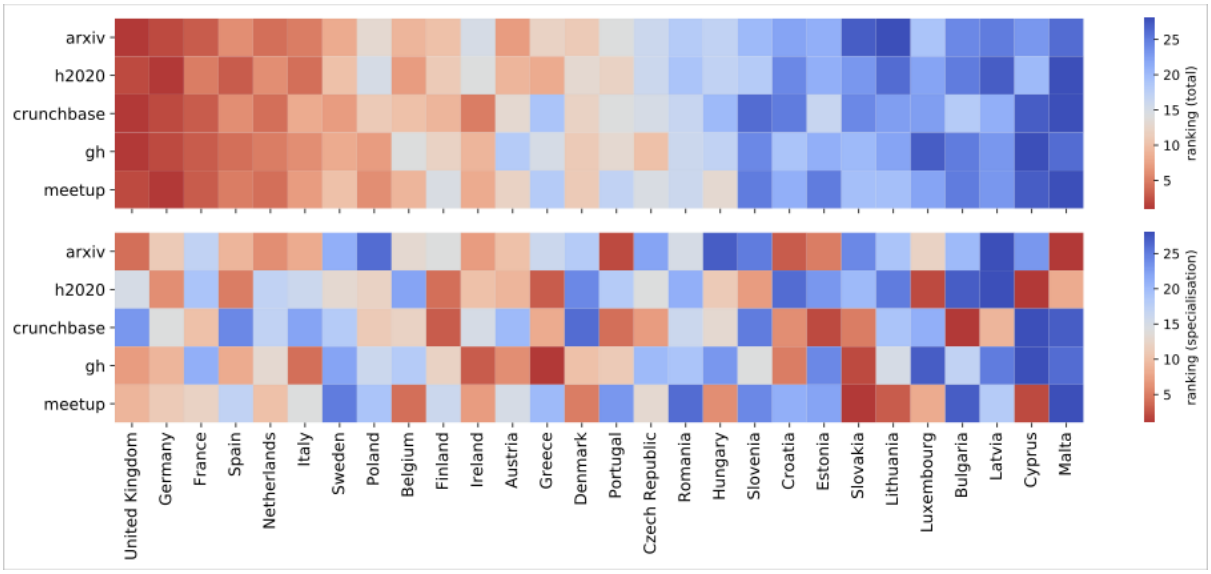


- What is the structure of different AI technology ecosystems inside the EU?

The heatmap below integrates all the dimensions of AI activity we have considered in the pilot and considers the ranking of countries in absolute terms and by specialisation. This kind of representation could help identify strengths and gaps in an AI technological ecosystem, as well as opportunities for new connections between EU countries. It could also be useful for identifying 'configurations' of AI ecosystems linked to differences in the structure of a country's technological innovation system and industrial structure, policy frameworks etc..

For example, Spain has moderate to high levels of specialisation in all areas except AI entrepreneurialism - perhaps this gap could be addressed through targeted interventions? The situation in France and Slovakia is almost the opposite, with high levels of AI entrepreneurialism in the face of low comparative advantage in research. We also identify some countries such as Portugal, Estonia or Croatia with apparently strong AI research bases but limited participation in H2020 AI projects. All this information, if made available with more detail, could help organisations find collaborators for their AI R&D projects, strengthening the EU research and innovation system.

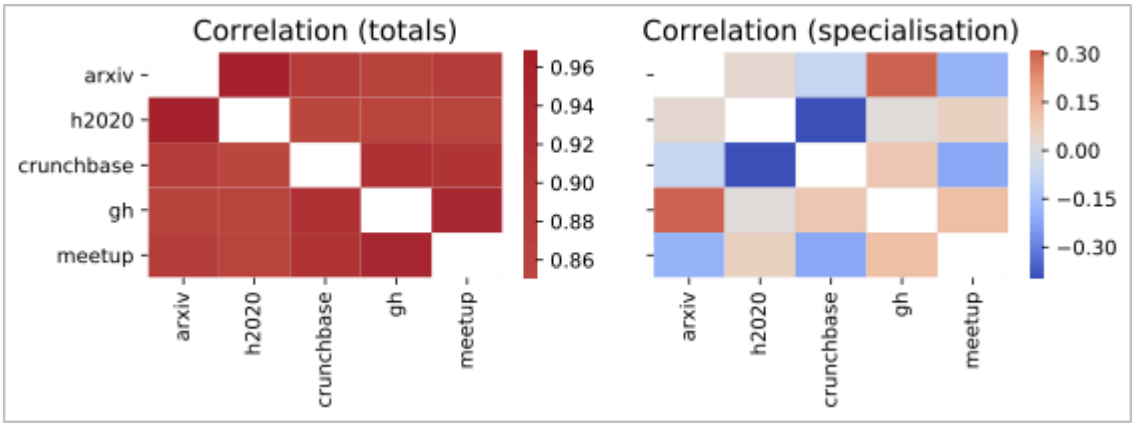
Figure 7



We also start exploring the idea of different models for AI technology ecosystem development in the correlation matrix below, where we consider the propensity for different activities to take place in (or experience an advantage in) the same countries.[5] It shows that the co-location of AI activities is stronger between those that are research oriented (research in arXiv and funding via H2020) and between those that are business/application oriented (CrunchBase and GitHub and, to a lesser degree, meetups).

Interestingly, we find a negative correlation between a country's relative advantage in AI H2020 funding and its strength in AI entrepreneurship. This result could be linked to the previously mentioned pattern where Eastern European countries display relatively low levels of participation in H2020 AI projects but relatively higher levels of entrepreneurship.

Figure 8



[5] The correlation uses the Spearman coefficient to reduce the impact of extreme, potentially noisy RAIs for lower-observation countries.

What is the regional distribution of AI R&D activity inside the EU and how does it compare to R&D activity in general, and to the distribution of the population?

We move on to consider the regional concentration of AI activity at the NUTS-2 level. In Figure 8 below, the maps in the first column represent the absolute levels of activity per data source in EU NUTS-2 regions, and the second column presents the cumulative share of AI activity by NUTS-2 regions ranked by their position in the distribution of activity in the data source compared with the share of all activity in the data source, and the share of population based on Eurostat.

Both sets of figures suggest high level of concentration of AI-related activity, with a small number of regions accounting for most of the activity. The cumulative curves show that, for all countries, AI-related activity is more concentrated than the population, suggesting the presence of agglomeration economies driving concentration of AI R&D, and also (if less starkly) than the overall levels of research/business/coding/ activity in each data source. This is consistent with the finding in Klinger et al (2018) that AI-related computer science research tends to be more geography concentrated than computer science in other application areas because its general-purpose nature benefits from co-location with a wider set of research bodies and industries.

We can also compare the propensity towards geographical concentration of activity across data sources by looking at the slope of the cumulative share curve. A steeper slope indicates that a more significant share of activity is accounted for by the larger regions. Here, we see that measures of activity related to research (arXiv and H2020 funding) tend to be less geographically concentrated than industrial measures (start-ups in CrunchBase and technology meetup activity), hinting at the potential role of universities and research funding in making the geography of innovation more inclusive, but also underscoring the risk that research takes place in locations where there are few opportunities for diffusion into industry. It is interesting to note high levels of concentration in AI-related GitHub activity, despite the fact that this is a platform hosting virtual communities. This could reflect concentration in the AI workforce, or the fact that many of the GitHub projects reflect local collaborations.

In Figure 9 we compare the Herfindahl indices of concentration by data source and country (recall that a higher score implies a higher level of geographical concentration of activity), with countries ranked by total levels of activity in the horizontal axis. We could think of this as a proxy for the extent to which AI activity is spatially equal - or unequal - inside each data source and EU member-state. The figure shows that in countries with lower levels of activity (in the right side of the scale), it tends to be highly concentrated in a single location. When we look at larger countries, we find some interesting differences: France, Finland and Ireland present high levels of spatial inequality in all dimensions of AI activity. The UK is relatively equal in research related activities, and unequal in business-related activities such as CrunchBase and Meetup. Germany is highly decentralised in all measures, suggesting the presence of several strong hubs of AI activity in the country.

Figure 8

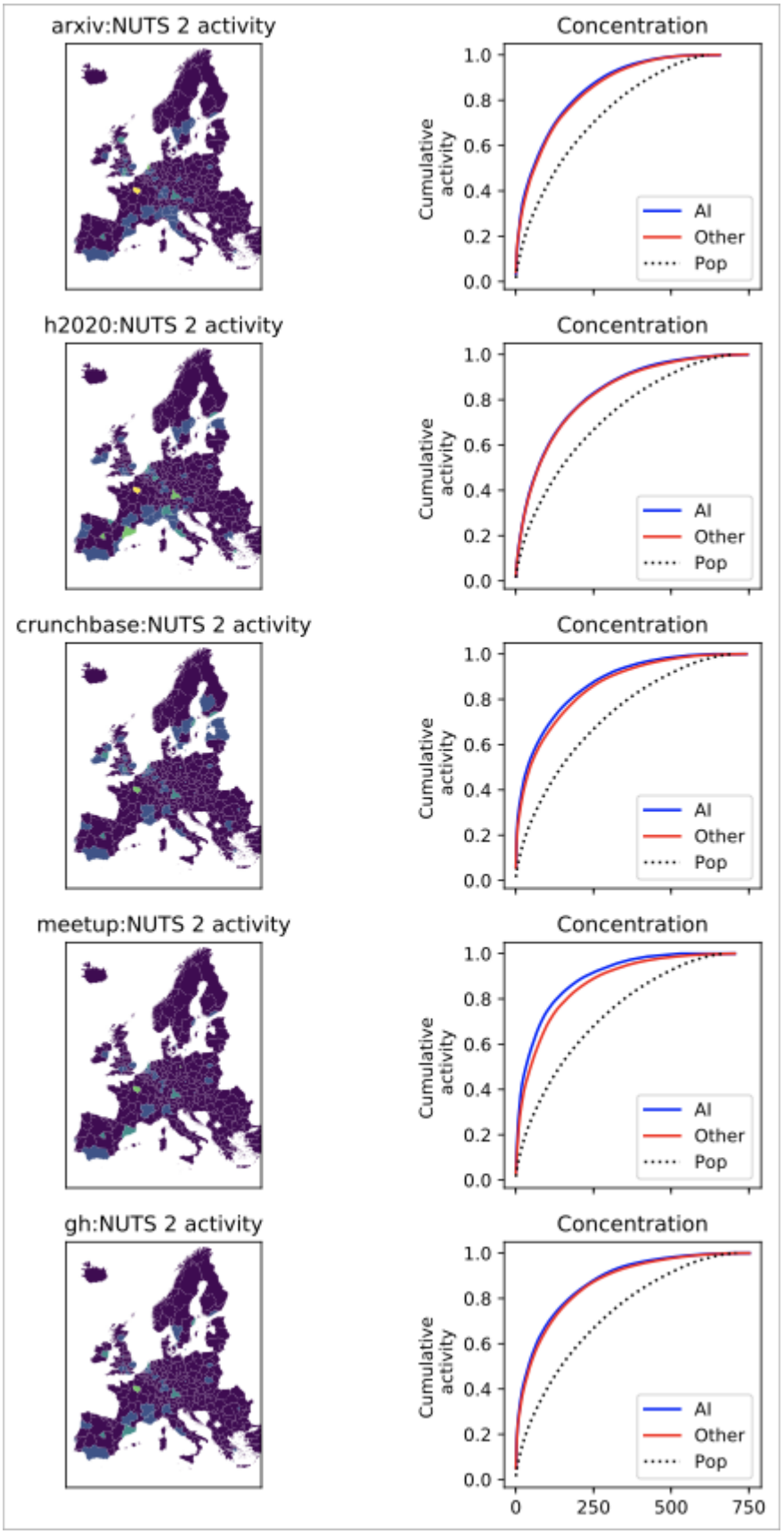
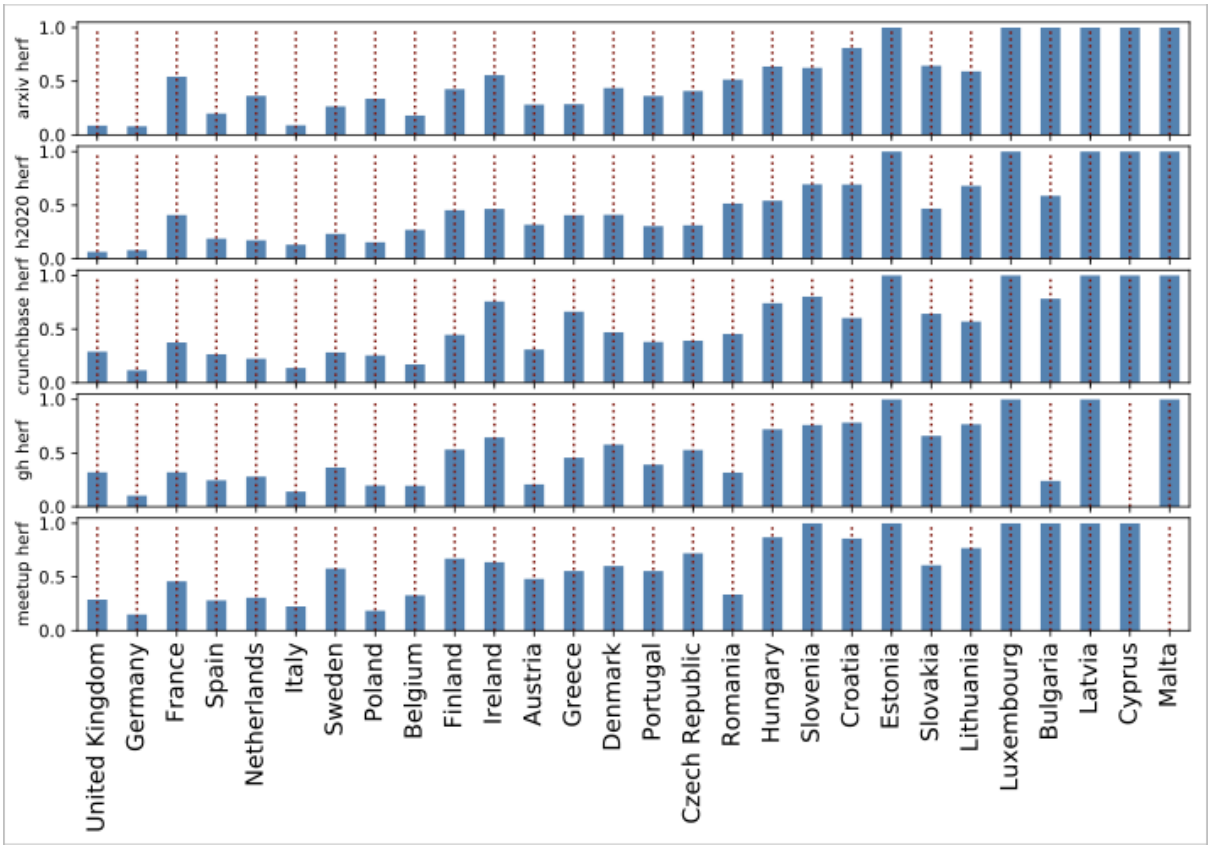

arxiv:NUTS 2 activity
Concentration
h2020:NUTS 2 activity
Concentration
crunchbase:NUTS 2 activity
Concentration
meetup:NUTS 2 activity
Concentration
gh:NUTS 2 activity
Concentration
Cumulative activity
AI
Other
Pop
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750

Figure 9



What is the sectoral focus of AI R&D and how does it diverge between EU member states?

We conclude by reporting the findings of an experimental analysis of the industrial focus of AI activity in different data sources and countries. Our question is as follows: 'what is the industrial orientation of AI activity in different countries and data sources'?

Answering this requires classifying all entities in our data into the industries for which they might be relevant based on the language they use in their descriptions.[6] We do this following a supervised machine learning model trained on the CrunchBase data, where companies have already been tagged with industry labels. This involves several steps:

1. The CrunchBase taxonomy consists of 46 industrial categories. In order to reduce the number of target labels to consider during the analysis we cluster the original labels in the taxonomy into a smaller set of industrial segments, based on their co-occurrence in companies.[7] This results in 25 categories that we label by concatenating the names of their constituent CrunchBase categories.
2. We convert the CrunchBase data into a labelled dataset where every company is labelled with the industry clusters present in it based on the analysis in #1.
3. We train a machine learning model of a company's industrial segment using the pre-processed text in company descriptions as features.[8] The figure below presents the performance of the

[6] This is not the only approach. We could for example try to determine the sector of companies involved in research collaborations.

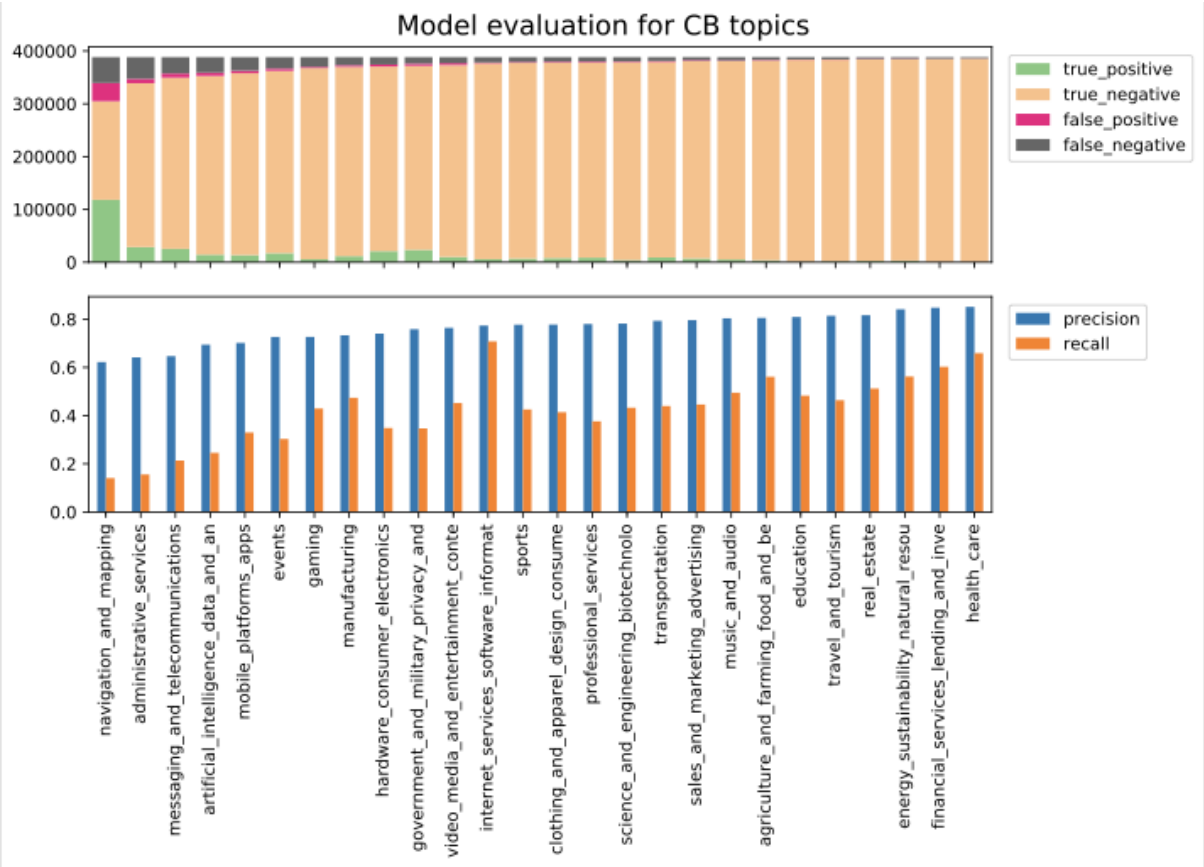
[7] We calculate the jaccard distance between categories and cluster them using a K-means algorithm.

[8] We use grid-search with 3-fold cross-validation with a logistic regression and random forests classifier tuning the parameters for class balance, type of regularisation and regularisation penalty. The best-performing model is the logistic regression with l1 regularisation and a C parameter of 0.1.

model in terms of the confusion matrix (first row) and precision-recall (second row) by industrial segment. We see significant variation in model performance across categories: performance is better in categories with tighter vocabularies such as health care, finance and energy.

4. We use the model to predict the industrial relevance of other entities in the AI-related dataset based on their textual description, pre-processed using the same model that we used in the CrunchBase data. We classify entities as relevant for an industry if its predicted probability for that segment is above 0.5

Figure 10



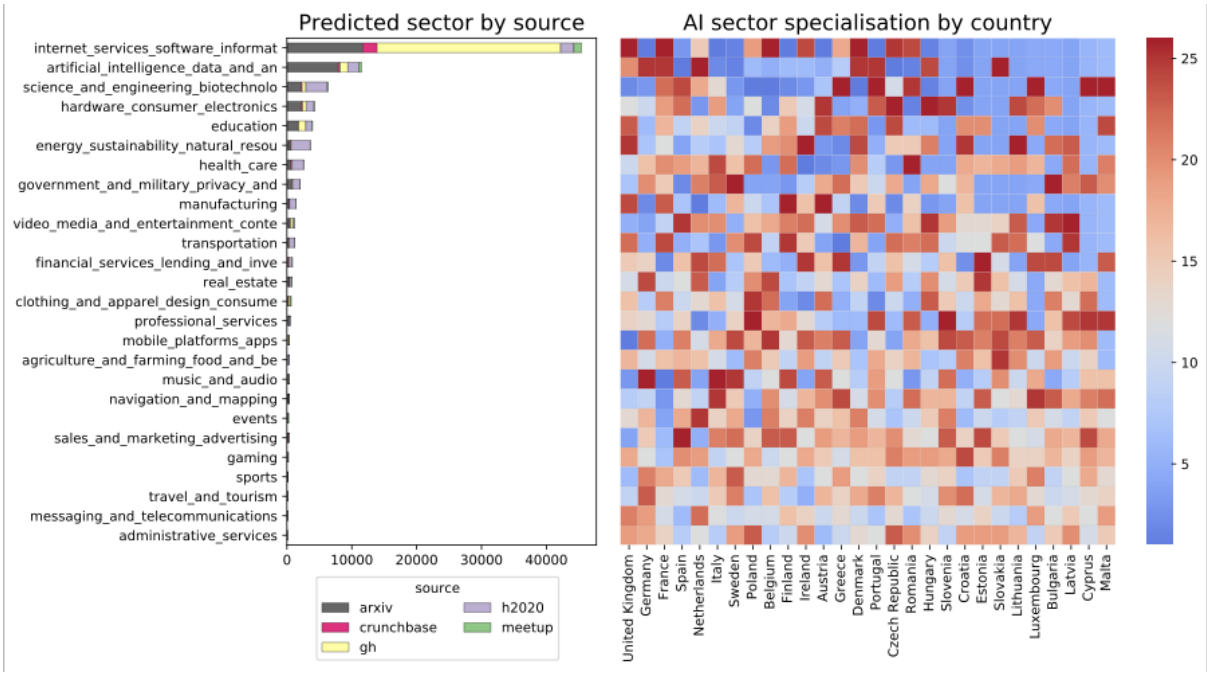
We advise special caution in the interpretation of the results below: as we previously mentioned, the text descriptions in the CrunchBase corpus are very short, impairing the performance of the model. Further, we are assuming that the relationship between keywords and labels that we find in CrunchBase holds for the other data. We consider options to address these shortcomings in the conclusion of the report.

Bearing all this in mind, in the figure below the left panel shows the total number of entities classified in each industrial segment by data source, and the right panel presents the national rank for each industrial segment in the portfolio of activity for each country, based on its RAI (warmer colours mean that an industrial segment ranks higher amongst all of a country's AI-related activities).

Perhaps unsurprisingly, this exploratory analysis suggests that most AI activity is related to "internet services / software", "data and artificial intelligence", "science, engineering and bio-technology" and "hardware and consumer electronics". We also see how relevant for different industries are different data sources. For example, GitHub has a stronger presence of internet-related activities, arXiv has more

activity in data and data and AI related activities, and H2020 has a stronger presence in Health, Science and Energy-Environment related domains.

The right-hand panel could be used to analyse the AI industrial specialisation profile in different countries: for example, we see that the UK is particularly strong in internet-related activities, education and energy, while Germany is strong on data, real-estate, mobile and music. Having said this, special caution is advised in the interpretation of these results - in addition to the potential issues with the supervised machine learning approach we have used, the right panel is combining data from multiple sources to calculate the RAIs by country and segment, and the RAIs for smaller categories and countries are probably noisy. A more disaggregated analysis would be desirable in the future. We do however believe that this analysis and representation of the data in this sub-section illustrate the potential for using data science methods to understand the industrial orientation of AI research in different EU countries.



4 Discussion and Conclusions

The findings above illustrate the potential for using novel data sources and analytical methods to generate highly relevant indicators about R&D activity in emerging technologies of interest for policymakers, such as AI. Further, the high granularity of the data makes it possible to analyse its spatial and sectoral distribution to generate measures of concentration relevant for spatial and industrial inclusion agendas (i.e. is an emerging technology only being developed in a small number of locations and sectors, or is its development more evenly distributed?). Finally, our consideration of multiple data sources capturing different dimensions of an AI technological innovation system could help identify gaps in those systems informing targeted interventions, as well as opportunities for collaboration between EU member states.

4.1 Validation and ongoing stakeholder engagement

We plan to triangulate our data with official sources and create ‘quality filters’ for data sources we are currently using where they might be concerns about garbage contributions and spam. For example, we

can exclude from the analysis pre-prints that cannot be matched with a peer-reviewed publication or conference proceeding, open source projects with no code contributions, or meetup tech groups with numbers of events or members below a certain threshold. An important consideration for visualisation will be whether to set these thresholds automatically (at the risk of excluding emergent activities) or let users select them (at the cost of increasing complexity in the visualisations).

We will also continue the validation of findings and engagement with key stakeholders in the EU including DG RTD, the JRC and the High-level expert group on AI. We also plan to use some of the thinking and emerging results from the analysis to publish a short evidence briefing calling for the use of novel data sources and methods to generate relevant, inclusive, trusted and open data to inform policies around emerging technologies in the EU.

4.2 General limitations

We have already touched on some of the limitations of our approach above. Several of them relate to implementation and can be addressed through additional data collection to increase the range of sources we are considering, and further enrichment of those we are already using. We also need to explore better ways to present the rich data that we are collecting through the pilot.

More fundamentally, an important limitation of our approach that was highlighted in the Knowledge Stakeholder Workshop is the fact that we rely on the users to generate an initial list of keywords to search for. Put simply, our approach can give users relevant information about emerging technologies that they are already aware of / interested in, but it is less useful for discovering new interesting trends to analyse. We believe it is possible to address this limitation of our analysis by combining this pilot with work taking place elsewhere in EURITO (we discuss this further below).

While our focus on producing simple indicators has the advantage of making our analysis easy to communicate, it removes much relevant information about technological innovation systems, such as for example the structure of collaboration networks between researchers and industry, and between countries and regions. Those analyses could perhaps be used to estimate additional indices of connectivity about EU and national technological innovation systems.

4.3 Considerations for scaling up

As mentioned, all the data sources that have used in the pilot are domain-agnostic (although with a potential bias towards digital activities). Our use of Clio, a flexible expanded keyword-based approach to identify relevant entities in the data reduces our dependence on sectoral taxonomies. None of the data we are currently using are country specific, which enables EU-wide or even global analyses of emerging technology trends.[9]

Some key considerations for scaling-up include:

- Collect new data about papers, patents and skills supply
- Enrich available sources with additional information to improve the quality of our queries and supervised machine learning analysis
- Integrate data into a single database

[9] Having said this, there might be some biases in country / region coverage by different sources if for example publication / code-sharing practices diverge, or platforms that we are considering have local competitors. We will consider all these issues during the validation stage of the project.

- Continue the development of Clio to address the limitations identified above
- Explore additional avenues to measure the industrial relevance of different entities in the data, for example using open corpora such as Wikipedia or existing taxonomies of industrial activity.
- Explore different strategies to visualise the data

4.3.1 Complementarities with other pilots

This pilot has strong complementarities with several others in the EURITO project:

- Pilot 3 (Technological Change) could be used to identify emerging technologies to query against Clio in order to generate technological ecosystem indicators, or to generate country or regional level measures of structural transformation.
- Pilot X (Inclusive innovation) could be used to measure the level of socio-demographic inclusion in emerging technological innovation systems based on the names of paper authors, inventors etc.
- Pilot Y Z (Advanced Funding Analytics, Mission Driven and Knowledge Flows) are using research output data from the Open AIRE platform which could be used to develop additional indicators of knowledge exchange from emerging technologies at the national / regional level. Further, some of the data on societal responses to research being collected in the Mission-driven innovation pilot could be used to generate indicators of societal perceptions about emerging technologies, an important factor shaping their development (Mazzucato, 2018).
- Pilot W (Standards) involves data that could be used to proxy the diffusion of emerging technologies into industry.

4.3.2 Tools and data sources

This pilot uses Python. All the analysis was performed in memory in a 16GB 2.4GhZ laptop and documented in JuPyteR notebooks. This repo (http://github.com/nestauk/eurito_pilot_1_emergent) contains draft versions of the notebook still to be refactored and documented.

5 References

Aghion, P., David, P. A., & Foray, D. (2009). Science, technology and innovation for economic growth: Linking policy research and practice in 'STIG Systems.' *Research Policy*, *38*(4), 681–693. https://doi.org/10.1016/j.respol.2009.01.016

Agrawal, A. K., Gans, J. S., & Goldfarb, A. (2018a). *Economic Policy for Artificial Intelligence* (Working Paper No. 24690). National Bureau of Economic Research. https://doi.org/10.3386/w24690

Agrawal, A. K., Gans, J. S., & Goldfarb, A. (2018b). *Exploring the impact of artificial intelligence: Prediction versus judgment*. National Bureau of Economic Research.

Agrawal, A., McHale, J., & Oettl, A. (2018). *Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth* (Working Paper No. 24541). National Bureau of Economic Research. https://doi.org/10.3386/w24541

AI Index. (2017). *The Artificial Intelligence Index: 2017 Annual Report*. Retrieved from http://cdn.aiindex.org/2017-report.pdf

AI Index. (n.d.). *AI Index*. 2018. Retrieved from https://aiindex.org/

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *ArXiv Preprint ArXiv:1708.05866*.

Bakhshi, H., & Mateos-Garcia, J. (2016). *New data for innovation policy* (Working Paper). London: Nesta.

Bergek, A., Jacobsson, S., Carlsson, B., Lindmark, S., & Rickne, A. (2008). Analyzing the functional dynamics of technological innovation systems: A scheme of analysis. *Research Policy*, *37*(3), 407–429.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Börner, K., Rouse, W. B., Trunfio, P., & Stanley, H. E. (2018). Forecasting innovations in science, technology, and education. *Proceedings of the National Academy of Sciences*, *115*(50), 12573–12581.

Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., … Evans, J. A. (2018). Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, *115*(50), 12630–12637.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). *The Impact of Artificial Intelligence on Innovation*. National Bureau of Economic Research.

Dalle, J.-M., den Besten, M., & Menon, C. (2017). Using Crunchbase for economic and managerial research.

Elsevier. (2018). *Artificial Intelligence: How knowledge is created, transferred, and used*. Elsevier.

European Commission. (2018). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe.* Brussels: European Commission.

Goldfarb, A., & Trefler, D. (2018). *AI and International Trade*. National Bureau of Economic Research.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hain, D. S., & Jurowetzki, R. (n.d.). Introduction to Predictive Modeling in Entrepreneurship and Innovation Studies.

Hicks, D. (2011). Structural change and industrial classification. *Structural Change and Economic Dynamics*, *22*(2), 93–105.

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575. https://doi.org/10.1073/pnas.0900943106

Joint Research Centre. (2018). *Artificial Intelligence: A European Perspective*. Seville: JRC.

Jurowetzki, R., & Hain, D. S. (2014). Mapping the (r-) evolution of technological fields–a semantic network approach. In *International Conference on Social Informatics* (pp. 359–383). Springer.

Klinger, J., Mateos-Garcia, J., & Stathoulopoulos, K. (2018). Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. *ArXiv Preprint ArXiv:1808.06355*.

Klinger, Joel, Mateos-Garcia, J., Stathoulopoulos, K., Tippett, C., Moeremans, R., & Morret, J. (2018). *Exploratory Report B: Toward the incorporation of Big data in the European Innovation Scoreboard*. Brussels: European Commission. Retrieved from https://ec.europa.eu/docsroom/documents/30362/attachments/1/translations/en/renditions/native

Korinek, A., & Stiglitz, J. E. (2017). *Artificial intelligence and its implications for income distribution and unemployment*. National Bureau of Economic Research.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Levy, F. (2018). Computers and populism: artificial intelligence, jobs, and politics in the near term. *Oxford Review of Economic Policy*, *34*(3), 393–417.

Mann, K., & Püttmann, L. (2017). Benign Effects of Automation: New Evidence from Patent Texts.

Mateos-Garcia, J. C. (2018). The Complex Economics of Artificial Intelligence. *Available at SSRN 3294552*.

Mazzucato, M. (2018). Mission-oriented research & innovation in the European Union. *Brussels: European Commission*.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N13-1090

Mulgan, G. (2017). *Big Mind: how collective intelligence can change our world*. Princeton University Press.

Nathan, M., & Rosso, A. (2015). Mapping digital businesses with big data: Some early findings from the UK. *Research Policy*, *44*(9), 1714–1733. https://doi.org/10.1016/j.respol.2015.01.008

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton University Press.

Siebert, M., Kohler, K., Scerri, A., & Tsatsaronis, G. (2018). *Technical Background and Methodology for the Elsevier's Artificial Intelligence Report*. Elsevier.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580.

Trajtenberg, M. (2018). *AI as the next GPT: a Political-Economy Perspective*. National Bureau of

Economic Research.

Pilot 2: Nowcasting Business Research & Development

Abstract:

This pilot explores the feasibility of several approaches to nowcasting R&D expenditure in the EU based on official Eurostat data on intramural Business Enterprise R&D expenditure (BERD), and the Business R&D Scoreboard - a dataset with detailed historical information from the top 2500 industrial investors on R&D from the EU and abroad based on their annual reports going back to 2004. We complement the wealth of existing research by considering both the feasibility of extending existing approaches and by exploring alternative models that address some shortcomings of existing approaches.

1 Introduction

1.1 Background/context

Although public R&D drives the generation of knowledge and talent needed by innovative enterprises, it is primarily through business investment that the full impacts of R&D can be practically realised. Business R&D integrates and transforms available knowledge into commercially viable technologies and innovations such as new products, processes and services that enable higher productivity, competitiveness, and efficiency.

Business Enterprise R&D (BERD) investment is a policy relevant, lagging, and expensive to collect innovation indicator. The ability to "nowcast" - predict an estimate of a hard economic variable of interest for the current reference period (e.g. month/quarter/year) - BERD investment both temporally and spatially would allow policymakers to act with more up to date and granular information.

Such a system would therefore utilise a combination of official (lagging and aggregate yet accurate, purpose built, and readily available for researchers to download) and unofficial (timely and granular but containing hidden biases and costly for researcher to collect) data sources to train predictive models.

Understanding the drivers of BERD has been a topic of research for many years. For example, findings from a synthesis of studies by Falk (2006) conducted in OECD countries found that business R&D intensity is positively associated with R&D tax incentives (regardless of specification and estimation techniques) as well as expenditures on R&D performed by universities.

The R&D expenditure of a firm is a complex variable linked to economic factors (nationally and globally), competition, the firm's performance, a firm's investment, mergers and acquisitions, news events, and so many other factors. Predicting R&D expenditure in the future is for this reason an exceedingly challenging goal. However predicting R&D expenditure in the present or very recent past - i.e. "nowcasting" - is a more achievable goal. Current expenditure is still dependent on all the factors mentioned above which are too numerous and immeasurable to be fully accounted for in a model but the modeller at least has access to knowledge of influential events such as whether the economy is in recession or not, disasters, mergers/acquisitions etc.

Nowcasting official statistics

Nowcasting official statistics is not a new area, there is a large history of literature already dedicated to the subject; however in the majority of cases the literature focuses on nowcasting variables such as

inflation, unemployment, and GDP which are required at a higher than annual frequency and have longer time series than are reliably available for R&D expenditure indicators (only 10-20 data points). The lower frequency and volume of data constrains us to a much smaller part of the literature which we have very briefly explored.

Dernis (2007) identifies three primary nowcasting methodologies:

1) **Transfer rates**
   Assume that the ratio of a variable currently know to the nowcasted variable is constant. This is incredibly easy to implement but typically performs poorly as it assume the ratio between the variables is stationary. Transfer rates have been used to nowcast patent data (Dernis 2007; Eurostat 2010).

2) **Trend models**
   Past trends in the nowcasted variable are extrapolated to create predictions. These typically require long time series as for short time series, long-term periodic components could be unobserved or short-term oscillations may appear as regular periodic behaviour but not exist in the long-term. Examples of these models are simple regression trend models, ARIMA (Autoregressive Integrated Moving Average), and GARCH (Generalised Autoregressive Conditional Heteroskedasticity) models.

3) **Econometric models**
   Models based on other economic variables such as number of researchers, GDP, value added etc. and probabilistic models. Castle, Hendry, and Kitov (2013) provide a survey of this area with examples including Mixed-data sampling models (MIDAS; Ghysels et al. 2007, Andreou et al. 2011); factor models (Boivin and Ng 2005); and Bridge equations. These methods are the most promising as they incorporate more of our prior knowledge into the model but are also the hardest to implement as they require both domain expertise and modelling skill.

Mouchart and Rombouts (2005) provide a general methodology for nowcasting from poor data (missing values and short time-series) using clustered panel data models with a case study in R&D variables. A classic panel data analysis would either pool all the country data together, account for country heterogeneity by country-specific intercepts (fixed effect model), or incorporate heterogeneity in the innovation term (random effects).

Mouchart and Rombouts instead incorporate clusters of countries specific to each coefficient to be estimated into a unique model, this has the effect of trading an increased bias (due to restricting coefficients to be fixed within clusters) for reduced variance (a lower number of parameters). Due to the non-stationarity of most macroeconomic series they estimate the model using first differences and argue from a structural-economic point of view that R&D data are not mutually independent even conditional on explanatory variables. This is because decisions depend on previous ones therefore an i.i.d. framework is undesirable and a minimum of dynamics are desirable (despite the short time span of the series making inference harder). These dynamics are incorporated by the introduction of 'rupture' indicators due to structural economic changes and are added where the residuals of a polynomial trend model exceed a given threshold. Clusters are constructed based on the distribution of estimates for each parameter of an OLS regression on each country with the quantiles of estimates defining the boundary between clusters. They then fit their group-wise pooling regression model with ruptures. This methodology is applied to nowcast "Total Government Budget Appropriations or Outlays for R&D" (GBOARD) considering growth in real GDP, general government net balance as a percentage of GDP, growth in total full time employment, and growth in employment as exogenous variables but only

selecting growth in real GDP for their final model. Their approach provides a reasonable framework for nowcasting R&D data yet suffers several shortcomings that can be addressed, such as using labelled data twice to form clusters/rupture indicators; not providing robust uncertainty estimates; and imposing hard restrictions on country effect sizes.

Other relevant work on nowcasting economic variables is that of Hernández et al. (2018) on the PREDICT dataset methodology. Statistics at a high level of detail in industry breakdown typically involve a 3 year lag, therefore to shorten the delay the authors objective is to nowcast several economic variables including BERD. The methodology followed here is a first differences linear regression with the possible addition of a first order autocorrelation term tested using the Breusch-Godfrey test (tests whether the residuals of a linear regression autocorrelate), three "impulse dummies" to capture shocks, and a step variable as some variables behaved differently after 2010.

Finally another piece of work by the JRC has focused exclusively on forecasting BERD at member state level using Eurostat data - Isella (2017). They use 27 predictors predominantly from national accounts and labour statistics (where data is at least one year fresher than BERD data) and use an elastic-net regression, a common method which allows regression on more predictors than there are data-points (27 predictors vs 10-15 observations). They report small relative error for most member states, with the exception of Poland, Romania, and baltic countries - they attribute this to either data quality issues or small economies being sensitive to the individual occurrences such as the opening/closing of a company with large R&D investments.

IRI Industrial R&D Scoreboard

Another body of work relates to the analysis of firm-level data available from the R&D scoreboard (the Scoreboard henceforth - http://iri.jrc.ec.europa.eu/scoreboard.html). The Scoreboard uses data stated by companies in their reports and audited accounts and comprises the world's top 2500 (EU top 1000) companies by R&D investments the companies have themselves made. The Scoreboard is released annually reporting data for the previous year with the first issue beginning in 2004 (albeit with fewer companies the further back one goes) which provides a certain degree of longitudinality to the data for firms that persist in the top companies. A companies' R&D expenditure is that reported by itself and its subsidiaries, therefore while a company may be reported to belong to a particular country, it is likely that the R&D activity was performed in multiple different countries. Therefore, to estimate territorial R&D expenditure it is necessary to apportion a companies' R&D activity to the countries it was likely to have occurred in.

Caro and Grablowitz (2008) come to much the same conclusion. The authors make a thorough study of the complementarity of BERD and Scoreboard methodologies and conclude that direct comparison is not appropriate as the two statistics have been designed for different uses - matching at the firm level is required to account for these differences.

One example of the differences is that Territorial statistics are complete and the Scoreboard covers only a small number of top companies; however in reality BERD is highly concentrated at company, country, and sector levels (Top 2500 account for ~90% global investment, top 100 ~53%, top 3 countries ~63%, and top 4 industries ~60%) - this is also due to the fact that the top 2500 companies' reports correspond to ~600,000 subsidiaries. This concentration in the Scoreboard has two opposing primary effects: a small number of companies can provide a fair estimation of BERD, and small deviations in a few companies can lead to large prediction errors.

Cozza (2010) overcome the fact that BERD microdata is not publicly available using private collaborations with BERD statistics producers in several countries to produce aggregate data from

BERD microdata that can be compared to the scoreboard directly - due to the untimely (BERD microdata are much less timely than the scoreboard) and private nature of this collaboration this approach is deemed inappropriate for this pilot.

An approach not relying on microdata was undertaken at the JRC by Gkotsis et al. (2017) which assigned R&D activity of Scoreboard companies to countries proportional to a companies' patent inventor location with the objective of nowcasting territorial R&D statistics using the scoreboard. The patent portfolio of companies was obtained from the EC-JRC/OECD COR&DIP database relating to the report by Daiko et al. (2017) which matches PATSTAT (The European Patent Office's worldwide proprietary patent database) to a proprietary list of company subsidiary data. This approach facilitates the linkage of the Scoreboard to territorial BERD statistics by both estimating the internal innovation of Scoreboard companies (and subsidiaries) and indicating the level of inward vs. outward activity within a territory (based on local companies' outward activity and foreign companies' R&D activity corresponding to patents by inventors resident in a territory). The authors then explore the relationship between the two datasets and find reasonable agreement between them - though this is only analysed through correlation coefficients therefore whilst long-term trends may be stable, an analysis of the agreement between fluctuations was not performed. Their analysis of cross-border R&D activity estimates that EU headquartered companies perform R&D at home >75% of the time, though there is significant heterogeneity with this number being as low as 30% for e.g. the UK. A key limitation of this work is its reliance on the COR&DIP database which is not timely (the first and only version was published in 2017 and corresponds to 2015 Scoreboard companies).

Patenting behaviour is likely to be highly heterogeneous across countries and industries, therefore by assigning R&D activity to countries purely based on patenting portfolios is an error prone exercise. Work that may complement that of the above was undertaken by Camerani et al. (2018) in order to understand the extent to which firms publish. They use the 2014 Scoreboard and the Bureau Van DIjk's (proprietary) ORBIS database to match Scoreboard companies and their subsidiaries (569,919 companies in total) to the author affiliation addresses of publications. They find that corporate publishing is widespread with 84% firms contributing to at least one publication between 2011 & 2015. This publication matching could augment the patent portfolio approach to help apportion R&D activity using a combination of these indicators. The key limitation of their approach however is that their methodology was not automated and required manual checking of hundreds of thousands of records - an approach that must be automated if we wish to perform a nowcasting exercise regularly and in a timely manner.

1.1.1 Opportunity

The Europe 2020 strategy is the EU's agenda for growth and jobs for the current decade. It emphasises smart, sustainable and inclusive growth as a way to overcome the structural weaknesses in Europe's economy, improve its competitiveness and productivity and underpin a sustainable social market economy.

It defines 8 goals over 5 thematic areas (Employment, R&D, Climate change and energy, education, and poverty and social exclusion) with the goal for R&D being that 3% of the EU's GDP should be invested in R&D.

With 2020 fast approaching and BERD investment being the primary component of R&D investment this illustrates the necessity of having more timely R&D indicators as currently it would be beyond 2022 before we know R&D expenditure in 2020.

This pilot has the opportunity to push existing research forward on two fronts. Firstly, to implement more advanced nowcasting frameworks for country (and perhaps regional) level R&D statistics that provide predictive accuracy, interpretability, and uncertainty estimation. Secondly, to incorporate more timely or varied indicators into the nowcasting models. FInally the potential to combine and augment existing work on linking the Scoreboard to territorial R&D statistics and to increase the automation behind these methodologies.

1.1.2 Flexibility of application domain

Whilst the focus of this pilot is around nowcasting R&D indicators, the models developed and data sources utilised should in principle be adaptable to the nowcasting of other policy relevant indicators/econometric variables of interest.

1.1.3 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

This pilot has a high potential to bring both new modelling approaches from the periphery of policy to its core. In addition to the new modelling approaches new data sources will be incorporated into a nowcasting framework.

1.1.4 Stakeholder engagement summary

The EURITO Knowledge Stakeholder Workshop provided an opportunity to identify important questions and considerations for this pilot, including bringing to our attention several existing similar projects.

The level of granularity at which nowcasting could be conducted emerged as a key point with increasing levels of granularity becoming increasingly less timely measures - interest was expressed in nowcasting chiefly at the country level but also at a regional and sectoral level.

1.2 Relevance to RITO criteria

1.2.1 Relevant

Business R&D is a highly policy relevant metric.

1.2.2 Inclusive

By including new innovation activities as predictors (e.g. open source software development, startup data etc.) we might be able to identify ‘hidden’ types of innovation investment that influence R&D activity.

1.2.3 Timely

This is the key target of this pilot. Business R&D statistics are typically available with >2 years lag.

1.2.4 Trusted

By exploring models within a Bayesian framework we ensure that we account for uncertainty in the model fit allowing us to include uncertainty directly within our predictions aiding robustness and interpretability.

Furthermore, these models are formulated in a hierarchical manner that balances model simplicity and expressability which is particularly important for this pilot where we have few observations across years and poor coverage for some countries.

We will consider models with different levels of interpretability, and cross- validate results to ensure robustness.

1.2.5 Open

All data used for the pilot stage except Crunchbase and the COR&DIP database are openly available with plans to keep any additional future indicators as open as possible.
COR&DIP database is available by request
(https://survey2018.oecd.org/Survey.aspx?s=03d8572cdd7f492aa1ed5350b7b9f044)
Code to collect this data and to run the models is openly available on github at
https://www.github.com/nesta_uk/rnd_eu

1.3 Research/policy questions

This pilot addresses the need for more timely R&D expenditure data. Current indicators are issued annually with country level intramural BERD expenditure issued with a lag of approximately 2 years (as of February 2019 data is available for 2016 for most countries and 2017 for a handful of countries) and the R&D scoreboard giving R&D expenditure data for the world's top 2500 (EU 1000) companies has a lag of approximately 1 year (as of February 2019, the latest scoreboard edition was the 2018 edition which gives data for 2017).

Multiple research questions/directions will be addressed in order to attempt to provide more timely indicators:

- In order to use the scoreboard data to nowcast R&D investment it is necessary to detect the organisational and geographical structure of companies, in order to correctly attribute R&D spending. For example, an automobile company may have its headquarters in Germany, but an R&D site in the UK. This becomes crucial if nowcasting is done in combination with placecasting but is important for nowcasting - R&D activity in a particular nation/region will be influenced by variables specific to that geography such as skills availability, tax incentives, and co-location of similar industries. Additionally, a company may have subsidiaries that perform R&D activities and these may also be in different countries to the headquarters.

- What modelling framework gives the ideal trade-off between predictive accuracy, reliable uncertainty estimation, and interpretability?

- What national or sub-national indicators predict R&D investment/activities? Is there a relationship between BERD and factors such as skills demand, start-up locations, open-source software development, tech meetups?

- What conditions facilitate R&D investment? Based on the criteria established by the previous question what regions have unfilled potential for R&D investment due to market misallocation or agglomerative factors?

2 Methodology

2.1 Data sources

We explore two R&D statistics data sources, official Eurostat territorial Business-Enterprise R&D expenditure data (available to download from their website or API) and the Scoreboard data (available separately each year as an Excel spreadsheet).

The R&D data we choose to predict are expenditures for R&D performed within the business enterprise sector (BERD) on the national territory during a given period, regardless of the source of funds. This data is obtained from the EUROSTAT database (https://ec.europa.eu/eurostat/cache/metadata/en/rd_esms.htm). We also obtain GVA data from the EUROSTAT database's data on annual national accounts (https://ec.europa.eu/eurostat/cache/metadata/en/nama10_esms.htm).

The distribution of companies across NUTS regions can be analysed using data licensed from Crunchbase. Crunchbase is a 'frontier' dataset (https://about.crunchbase.com/about-us/) that is increasingly being used to explore the digital economy. Founded in 2007, Crunchbase initially tracked firms that appeared on the TechCrunch industry news site. It has since evolved into a wide-reaching, firm-level crowdsourced dataset with rich information on technology-oriented firms, founders, employees, investors and investments. The dataset is regularly updated and has near-global coverage (containing firms in over 200 countries, although country-level coverage is uneven), making it a valuable source of insight for large-scale analyses of the digital economy.

Meetup.com data can be used to analyse the distribution of Science, Technology, and Innovation related community meetup groups, and is available through a RESTful API (https://www.meetup.com/meetup_api/)

Software development activity of Scoreboard companies is analysed using public GitHub repositories. Github is a service used by software developers to host code in repositories (which can either be public or private). Users have profiles that include personal information and project information, which may be linked to firm Github accounts. Github data is available from GHTorrent (http://ghtorrent.org/) in bulk or can be queried through the Github REST API (https://developer.github.com/).

Patents are obtained via. the COR&DIP database available by request from the OECD (https://survey2018.oecd.org/Survey.aspx?s=03d8572cdd7f492aa1ed5350b7b9f044), this contains the patent portfolios of 2015 Scoreboard companies.

Publications for Scoreboard companies are searched and obtained from the Microsoft Academic Graph (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/) API.

2.2 Methods

Scoreboard normalisation

An attempt was made to normalise the scoreboard data across years (censoring for countries that do not appear in all years); however this presented several difficulties. Firstly, the schema of the scoreboard changes annually as well as several variables changing year to year such as companies changing names, and the industrial classification codes used changing multiple times with no documented remapping. To overcome this we attempted to fuzzy match companies across years and then use a combination of fuzzy matching approaches to remap sectors. Despite this, the quality of

the harmonised dataset is relatively poor with many missing values present in the scoreboard and many other unattributable discontinuities due to factors such as human error and changing financial reporting standard.

If a cleaner dataset could be achieved this would assist in understanding firm level R&D investment behaviour longitudinally and would also provide more indicators to use for direct nowcasting.

Furthermore, one could bypass a firm-level matching by using statistical learning methods to estimate the cross-border R&D flow matrix of countries (How much of country *i* R&D occurs in country *j*) using the historic time-series of territorial statistics and the longitudinal Scoreboard time series (aggregated at country level). This would provide a method of using the scoreboard to nowcast territorial BERD data.

Finding scoreboard companies in other datasets

Software	**Automobiles and Parts**
SAP	Volkswagen
Amadeus	Daimler
Ubisoft Entertainment	Robert Bosch
Dassault Systemes	BMW
Atlassian Corporation	Fiat Chrysler Automobiles
Media	**Pharmaceuticals and Biotechnology**
RELX group	Astrazeneca
Sky	Sanofi
Technicolor	Bayer
Pearson	Glaxosmithkline
Daily Mail And General Trust	Boehringer Sohn

Table 1: Top 5 companies from a selection of sectors for the IRI EU industrial R&D investment scoreboard.

The top 5 companies from 4 industrial sectors (table 1) were chosen to assess the feasibility of using patent, publication, and software development data to allocate a companies activity to a given country.

Companies were identified within Github and MAG by semi-automated search, and patents matched to companies via. COR&DIP.

An analysis of the coverage and timeliness of these data-sources are presented in section 3.1.

Nowcasting models

We use a Bayesian framework to explore nowcasting BERD investment. Compared to the more typical Frequentist view of an events probability as the limit of its relative frequency in a large number of trials, Bayesian statistics interprets a probability as expressing a degree of belief in an event, which can change as new information is gathered. This interpretation is an incredibly powerful tool for performing statistical inference as it allows us to update the probability for a hypothesis as more evidence becomes available, and crucially gives us a full distribution across the hypothesis space - where frequentist statistics may give a point estimate for a quantity of interest (such as current BERD investment), a Bayesian approach gives us a distribution of estimates that becomes increasingly more certain with more data. Furthermore, Bayesian inference allows the incorporation of prior knowledge of a system which aids inference particularly in small datasets. For an introduction to Bayesian Data Analysis see either "Doing Bayesian Data Analysis: A tutorial with R and BUGS" Kruschke (2011) or "Bayesian Data Analysis" Gelman et al. (2013).

Put alternatively, where Frequentist statistics opts to use statistical tests to determine whether a null hypothesis (e.g. two groups are the same, or a parameter does not differ from zero) which involves several subjective decisions such as the statistical test to use, null hypothesis and significance level leading to the p-value fallacy, Goodman (1999), with Colquhoun (2014) showing that using p=0.05 (the de facto significance level) results in making the wrong decision 30% of the time. Bayesian statistics instead concerns itself with estimating how different two groups are, including uncertainty associated with that difference which includes epistemic uncertainty (uncertainty due to lack of data) and aleatory uncertainty (stochasticity of the system).

The use of Bayesian inference in nowcasting and econometrics is not new with a rich history going as far back as the 1960's - see Qin (1996) for an overview of the first 20 years of Bayesian econometrics. Whilst the computational and mathematical difficulty of performing Bayesian inference compared to frequentist inference has limited the full adoption of Bayesian techniques, this barrier has been steadily decreasing with the rapid progress in computation capabilities from the 1980's, widespread adoption of Markov Chain Monte Carlo (MCMC) techniques from Physics to statistics in the 1990's, improved MCMC sampling schemes of the 2000's and early 2010's to the recent rise of open source software and rich ecosystem of probabilistic programming languages. This rise of probabilistic programming languages such as PyMC3, Stan, and Edward which allow the specification and composability of probabilistic models along with efficient and general inference algorithms to fit the models have provided a new opportunity to efficiently develop and fit more complex Bayesian models, combine their predictions through techniques such as Bayesian model averaging, and even incorporate elements of deep learning into the models.

Here we begin to explore the possibilities such models offer this pilot using PYMC3 - Salvatier et al. (2016).

We try and predict BERD investment according to two transformations: on a log scale (*log_BERD*) to account for vastly varying country sizes and using log first differences (*log_D_BERD*) to avoid omitted variable bias (by cancelling out time-independent unobserved variables). Many EUROSTAT

indicators have been collected but time to fully analyse these has not yet been found, therefore for now we have chosen to use GVA as the sole exogenous variable.

For both *log_BERD* and *log_D_BERD* we fit various Bayesian methods, summarised in table 2 below with the relevant outputs of these models presented in section 3.2

log_BERD	***log_D_BERD***
Hierarchical linear model with partially-pooled country intercepts.	Hierarchical linear model with partially-pooled country slopes.
Hierarchical linear model with partially-pooled country intercepts and slopes.	Marginal Gaussian process trend model with long-term, periodic, and shock components

Table 2: Models used to nowcast territorial BERD

3 Results

3.1 Geocoding scoreboard

Preliminary studies have explored the R&D expenditure of the top five companies (by absolute R&D spending) in the automotive, pharmaceutical, media and software industries in the EU since 2006.

The trends and stability of R&D intensity over time within these companies varies greatly between sector (Figure 1). The pharmaceutical organisations spend the highest percentage, with the widest distribution, whereas the automobile industry spends considerably less across time, with a much tighter distribution across companies.

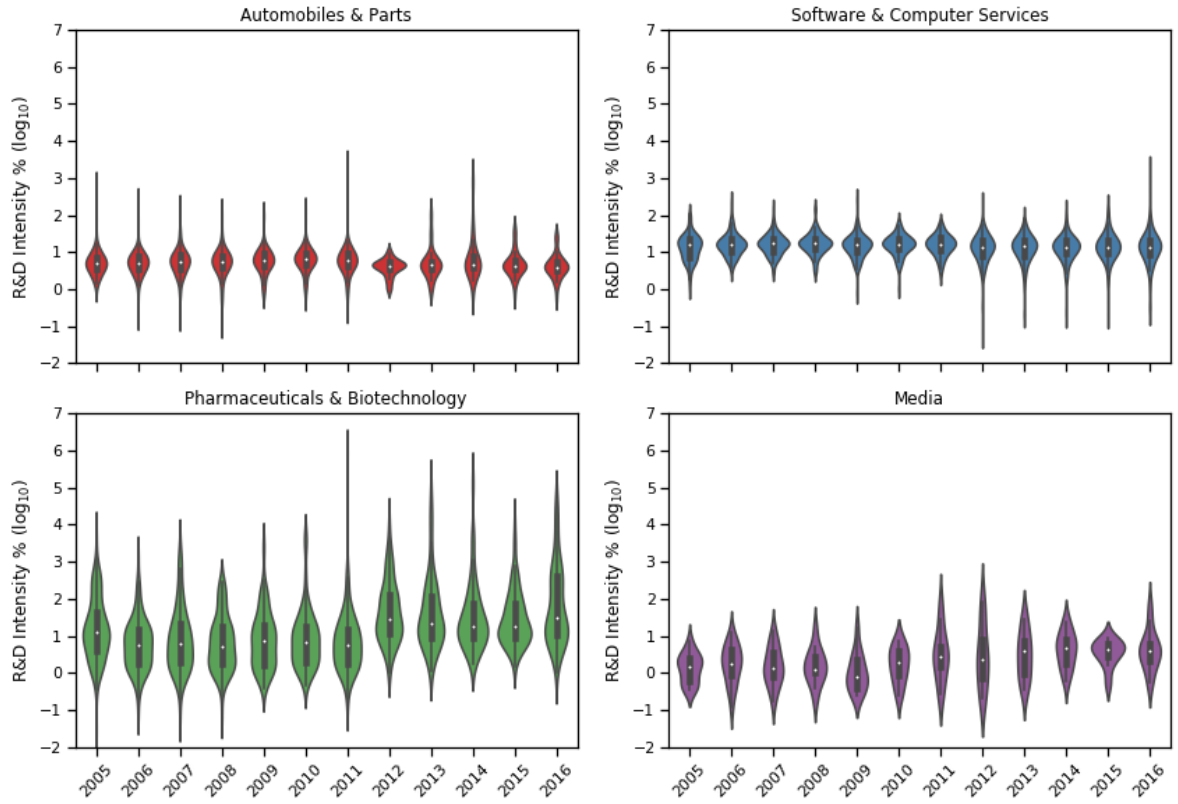


Figure 1: R&D intensity by industrial sector.

Github

Figure 2 shows the number of commits over time of the 20 companies, and their share of the number of annual commits. Atlassian and Sky were early adopters of Github , with their use of the platform continuing to grow but their share of the number of commits declining as other companies joined the platform.

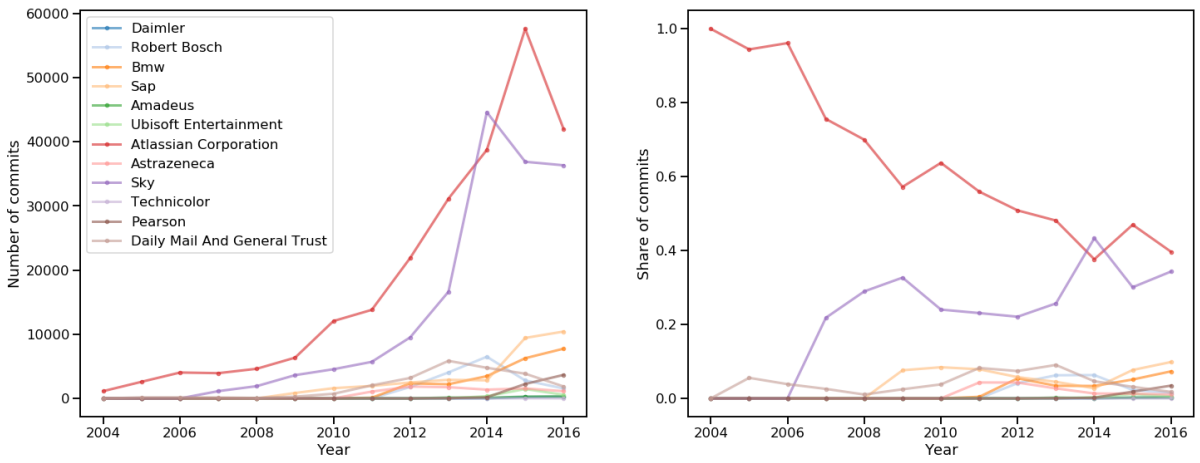


Figure 2: Github commits by companies over time.

Whilst the companies that have a presence on GitHub (many did not) showed a general upward exponential trend in activity (based on code commits), no discernable relationship between GitHub activity and R&D spending was observed. For example, Figure 3 displays the relationship between number of commits and R&D intensity for each company with the main pattern being the growth in the platform (constant R&D intensity, increasing number of commits over time).

A major difficulty in using Github data to measure activity over time is that normalisation is incredibly difficult - in order to assess whether a future rise in a companies Github activity corresponds to a rise in their R&D activity we need to control for growth in the platform in general as well as the adoption of the platform within the company. Furthermore, it is unclear how to best measure levels of activity as this could be done by number of commits, lines of code, number of active repositories and any other number of metrics each with their own biases.

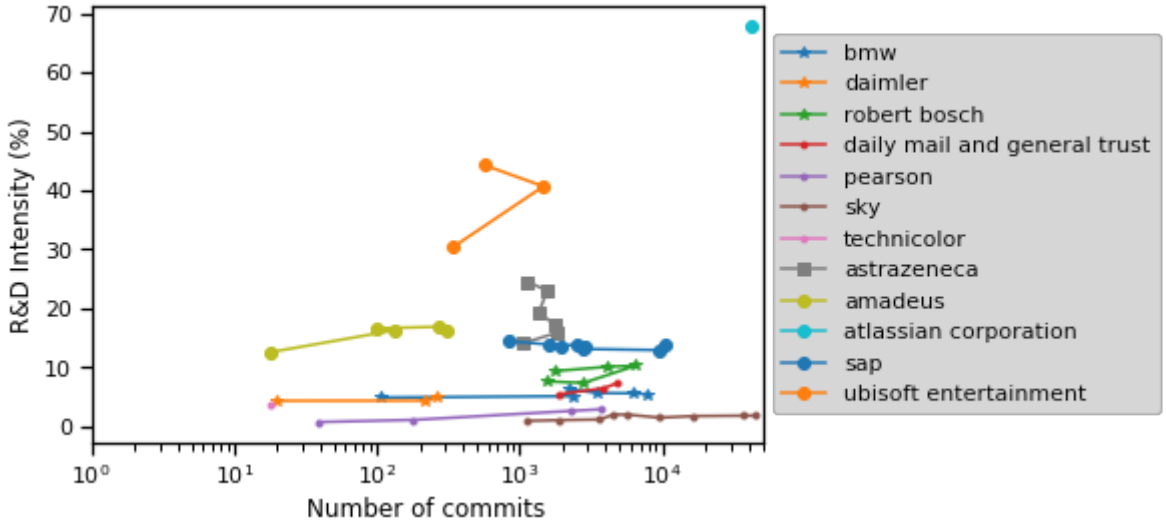


Figure 3: R&D intensity of companies versus Github commits. Note: Atlassian has an extensive commit history but was a recent newcomer to the scoreboard and therefore R&D intensity is only available for one year.

Publications

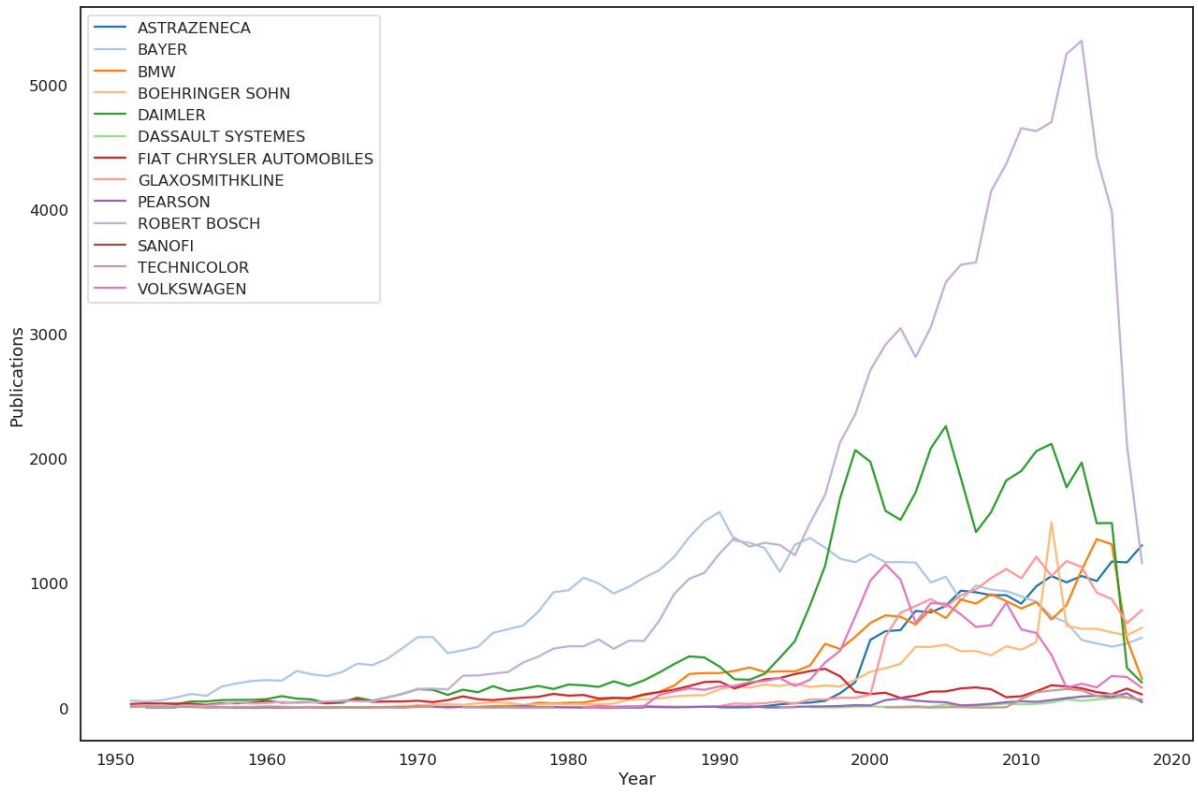


Figure 4: Publication count for top 20 companies.

The publication activity of companies over time (Figure 4) obtained from MAG shows good historic coverage for companies that engage in academic publishing; however the lag between R&D investment and publishing is unknown but the drop in publications for a number of companies in the last few years indicates that it is both significant and significantly variable.

Figure 5 shows the relationship between publication count and R&D intensity, there does not appear to be any meaningful relationship between the two. There are several potential methods of measuring activity value beyond raw publication count such as mean citation count, number of highly cited papers, average impact factor, and diversity of research. Different sectors will also have different behaviours making cross-sectoral comparisons difficult. How to measure the value of publications is in itself a policy issue with the advanced R&I funding analytics pilot attempting to address this.

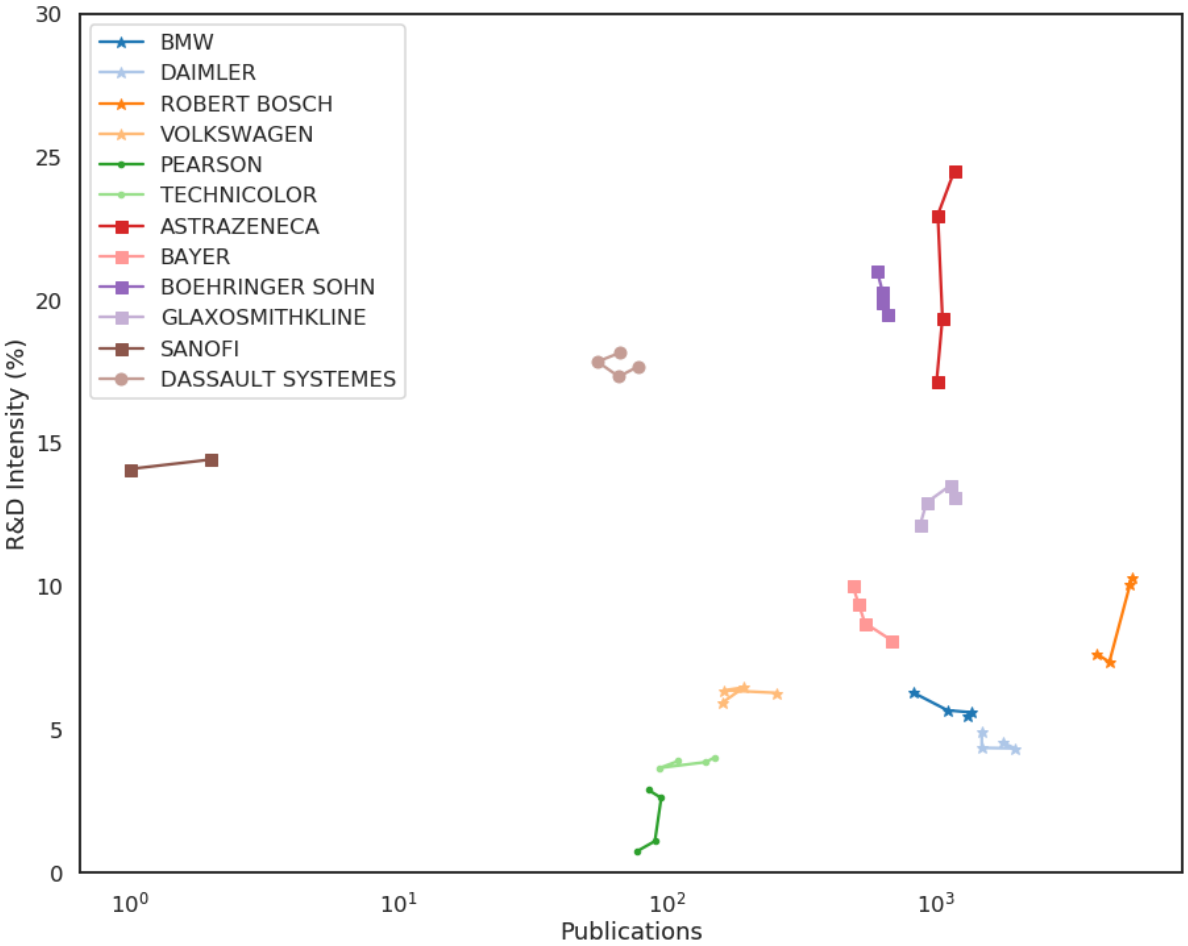


Figure 5: Publication count vs. R&D intensity.

Patents

Patents also have a long historic coverage, and all companies in the subset that were explored had engaged in patenting, though this may change for smaller organisations.

Figures 6 & 7 show the relationship between the number of patents filed by companies in a given year against R&D expenditure and R&D intensity respectively. It initially appears as log-patents and log R&D expenditure are linearly related but when stratified by sector this relationship is mostly flat apart.

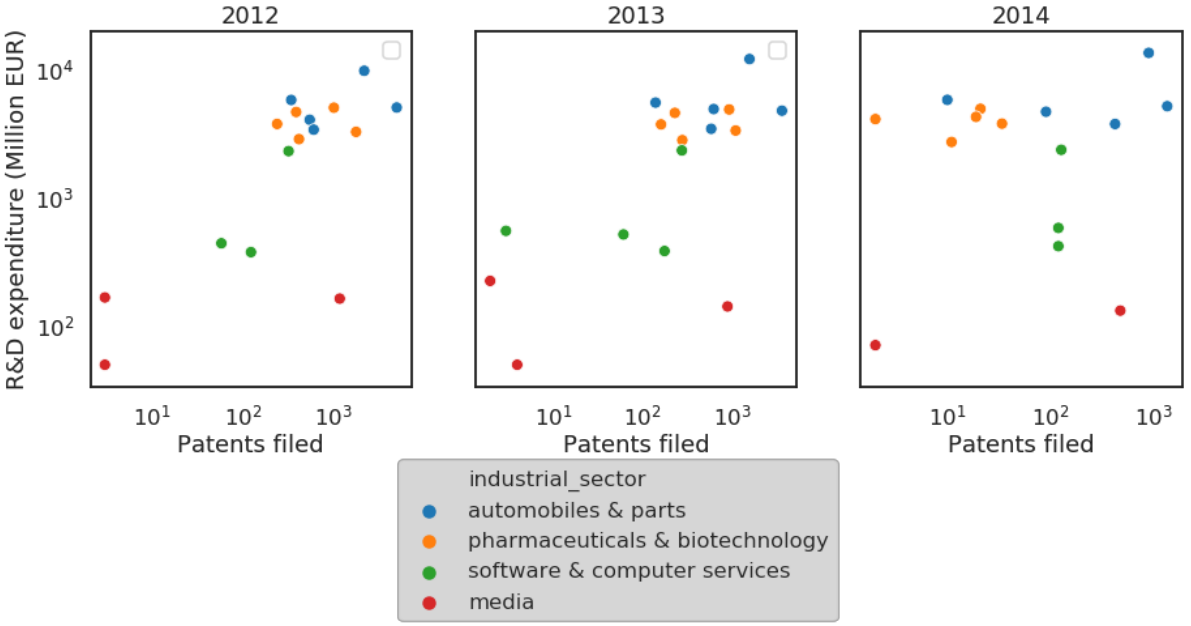


Figure 6: Log-log plot of patents against R&D expenditure.

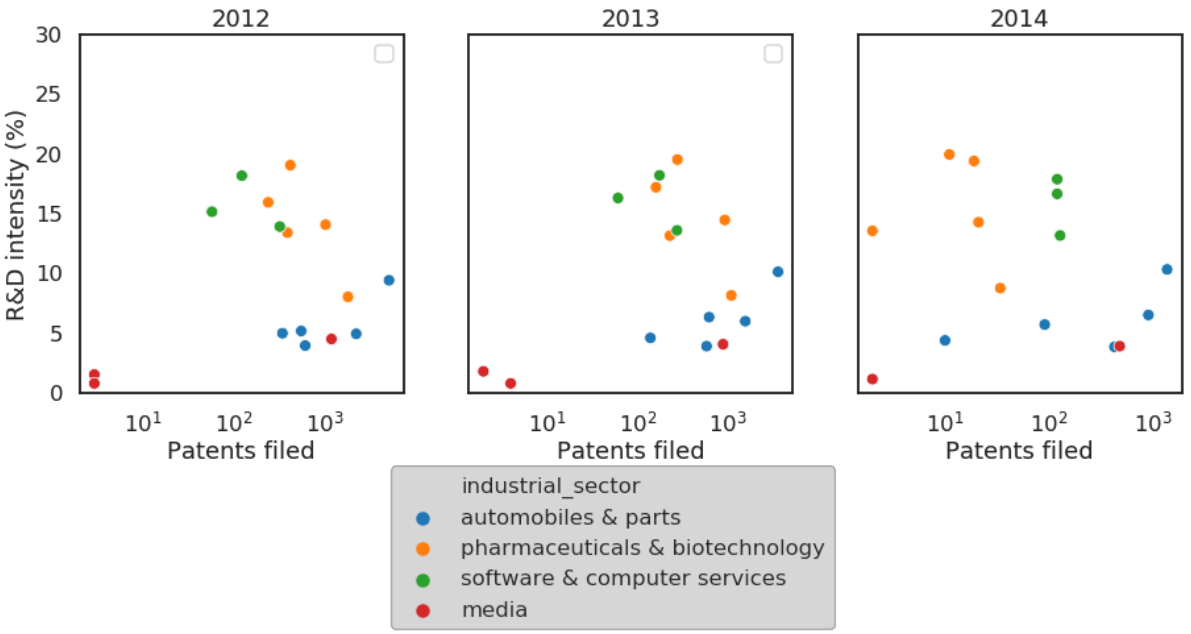


Figure 7: Plot of log-patents against R&D intensity.

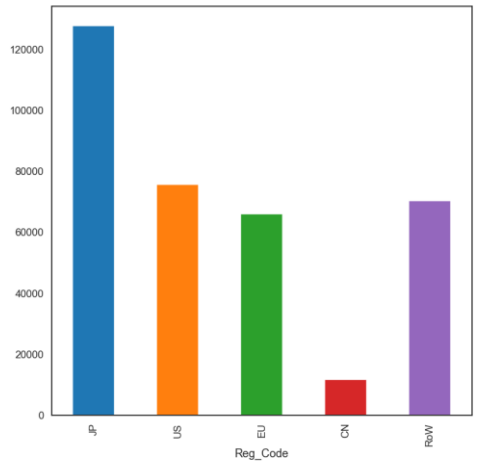


Figure 8: Patents filed by company region (Japan, US, EU, China, Rest of world)

Figure 8 shows the number of patents filed in different regions. Japanese companies are much more likely to patent than other companies with Chinese companies patenting far less than anywhere else.

As with publications and Github activity there are several potential methods of measuring activity value such as total number of patents, number of patent families, patent citations.

Geographic distribution of startups and meetup groups

Finding indicators available on a finer geographic scale would allow a better understanding of the R&D ecosystem of a country. We briefly explore the spatial distribution of Technology meetup.com groups and the location of startups in Crunchbase.

Meetup.com data could provide valuable insights into how different communities (that may correspond to different industrial sectors) interact and locate.

Crunchbase can provide insights into where new companies are locating. Given reliably geocoded R&D activity it would be interesting to measure the extent to which startups and R&D activities co-locate and which comes first. Better understanding this would allow the identification of areas which have all the requisite components to become a major startup/R&D hub but lack investment. This would most likely require access to BERD microdata as we have not formulated a viable method of attributing R&D activity on a regional level.

Crunchbase would also require verification for each country using something equivalent to the UK's IDBR (Inter-departmental Business Register) and census population data to estimate Crunchbases' coverage bias of the EU.

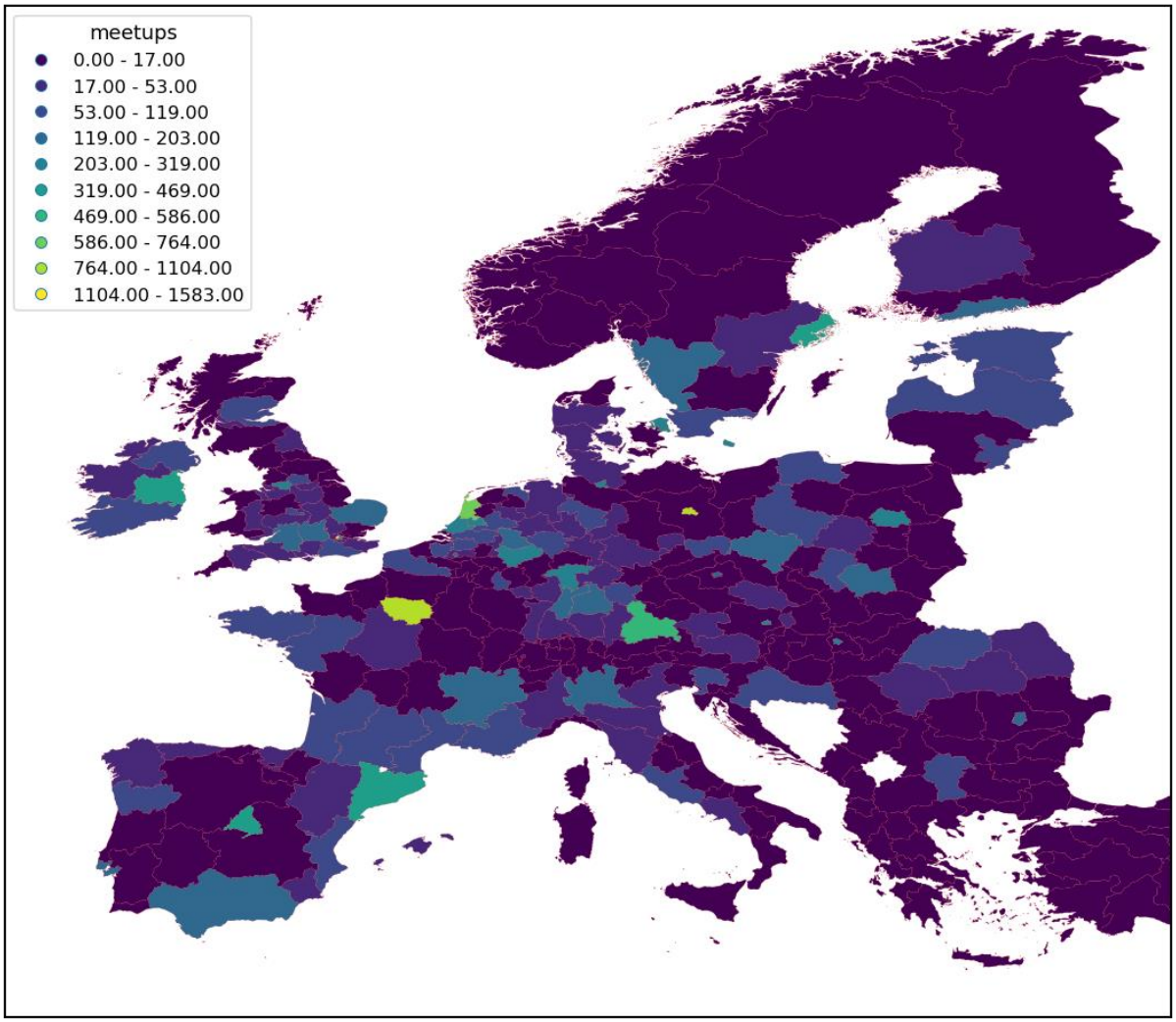


Figure 9: Choropleth map of tech groups from meetup.com.

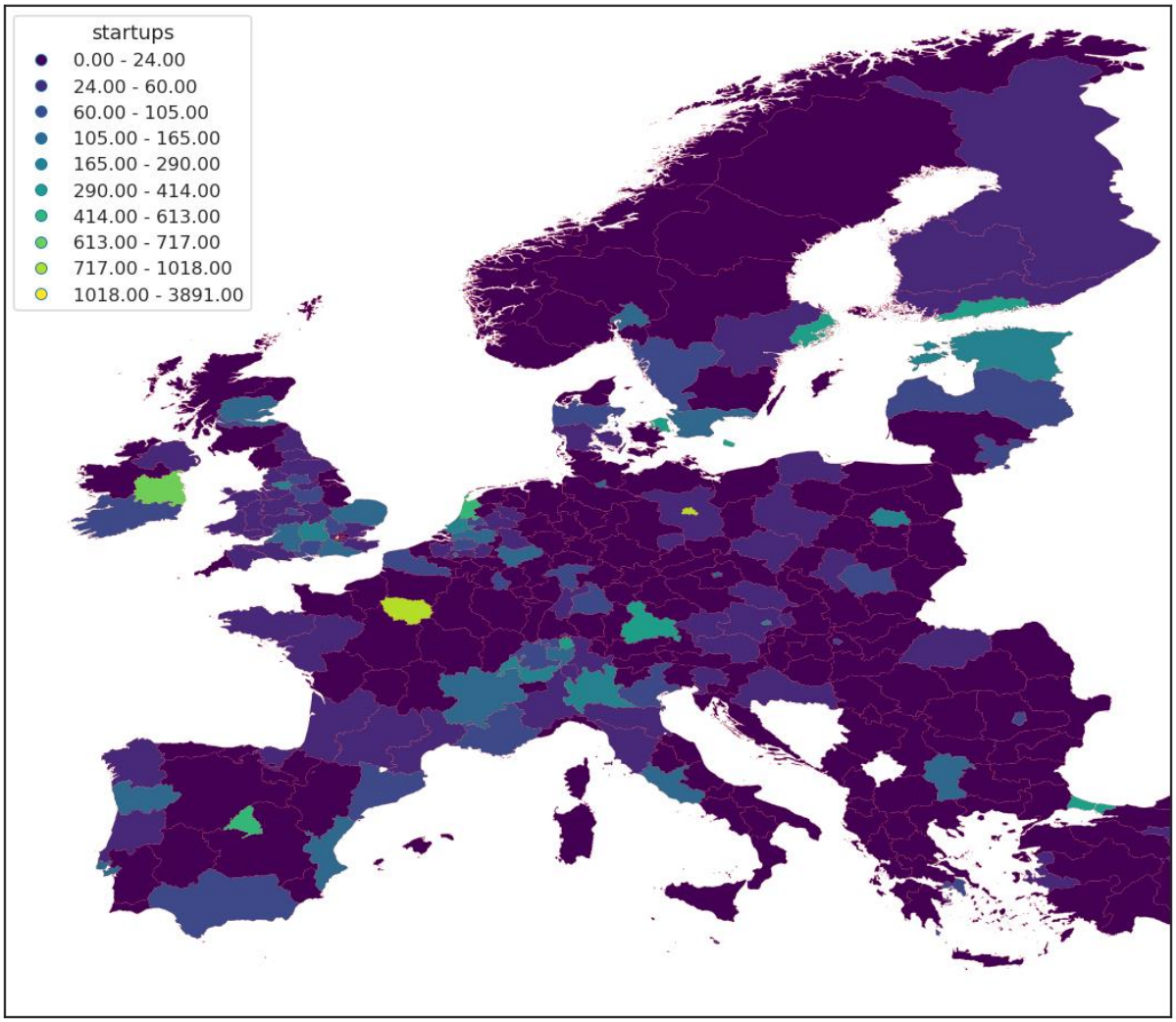


Figure 10: Choropleth map of startup locations from Crunchbase.

Both Meetup.com and Crunchbase data predictably concentrate around capital/industrial cities of Western European nations (Figures 9 & 10) meaning that using these datasets to perform placecasting would likely yield unreliable results for most regions. However, coverage in both datasets existed across most EU countries, suggesting that they may be useful for making predictions with variables at the national level. A future consideration is whether these datasets are appropriate for a longitudinal analysis as growth in these datasets is likely to be too highly correlated with growth in the platforms making it impossible to separate platform growth in a region with "knowledge growth".

3.2 Nowcasting BERD

Hierarchical linear models

First we employ a hierarchical (multilevel) linear regression model (neglecting temporality) using only log GVA as a predictor. A hierarchical model is where we assume that the regression coefficients of each country come from a common distribution in order to share information between groups, i.e. outlying countries that may have fewer observations (e.g. BA & MK) are shrunk towards the group mean - this is especially useful when we have little data.

The plot below shows the predictions for each country based on the first two models in the first columns of table 2. The model which included varying slopes was significantly better (verified using the WAIC score) as evidenced in the plot, this is because it allows a country specific response to the predictors whilst the partial pooling shrinking any outliers towards the group mean limits model overfitting due to this increased number of parameters. The predictions are not great but show a degree of promise using only predictor and for such short time series.

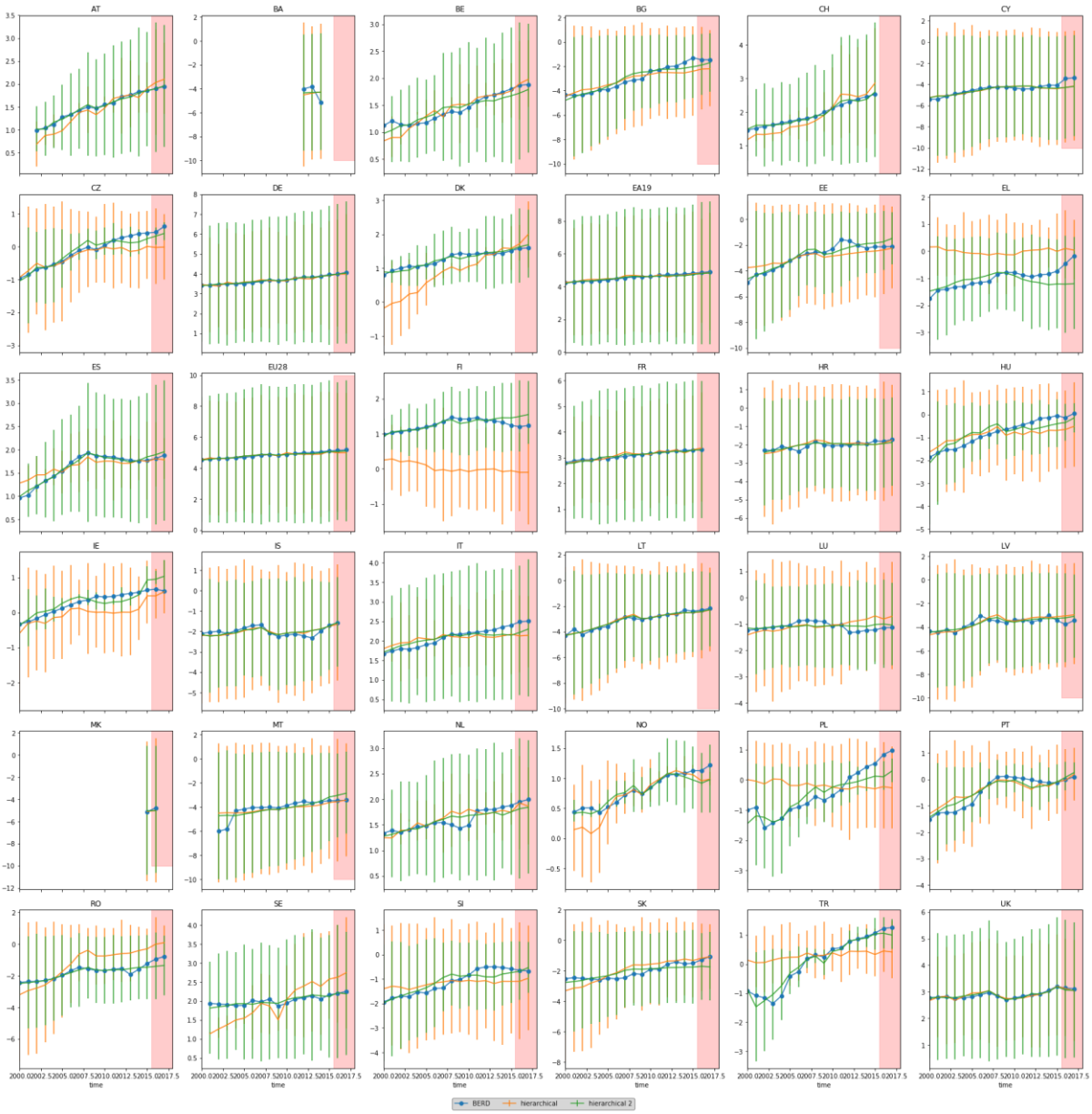


Figure 11: Centred log BERD predictions over time for two hierarchical linear models (uncertainty given by 50% Bayesian Credible Intervals)

Next we test a similar approach but using first log-differences of BERD and GVA to avoid the omitted variable problem where there may be some variables not included in the model that can be cancelled out by taking differences.

We employ a hierarchical linear regression model as before and show the results in figure 12. Unfortunately, the first log-difference approach yields lacklustre predictive capabilities. The posterior

predictive mean appears good for more developed regions e.g. UK, Norway but much less stable for smaller economies possibly due to firm-level fluctuations having large relative effects/data quality issues/chaotic behaviour. The predictions for each country are shown below. Though the model uncertainty bounds do capture most of the fluctuations, there is obviously finer structure that the model is missing that requires more (and more sophisticated) indicators, as well as the addition of impulse variables. These could be added empirically or using a more robust approach such as a stochastic volatility model or Gaussian processes.

Figure 12: Centred first log-difference BERD predictions over time for a hierarchical linear model (uncertainty given by 50% Bayesian Credible Intervals)

Gaussian Process models

Next we consider a Gaussian process trend model for the first log-difference transformed data. This does not use covariates such as GVA but is just modelling the trends throughout time. We do this to try and decompose the time series into multiple different components:

1) A long-term trend
2) A periodic term
3) Very short-term shocks

We set priors on each of these components to reflect the typical length scales and magnitudes we believe these components to have.

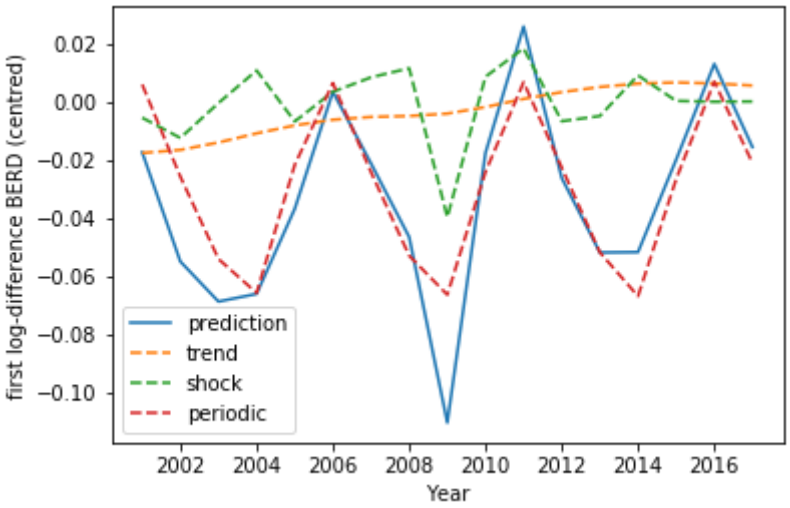


Figure 13: Prediction and trend components from the Gaussian Process model for Norway, UK, France, Germany, Belgium, Netherlands, Finland, and Sweden.

Fitting the model for a subset (there are computational efficiency issues for us to overcome as Gaussian processes have cubic complexity with the number of datapoints) of more R&D intensive countries (Figure 13) we see evidence of a periodic term with period of almost exactly 5 years, perhaps indicating the timescale over which firms investment cycles occur or boom-bust cycle of the economy. It is important to note that due to the short nature of the time-series we cannot draw any definite conclusions. We observe a large shock around 2008 corresponding to the financial crash. Finally, we see a gradual long term trend for BERD expenditure to increase - as we are modelling in log first differences this corresponds to an increase in the rate at which firms are investing in BERD!

The predictions against the data are shown in figure 14 below along with out of sample predictions for 2016 & 2017.

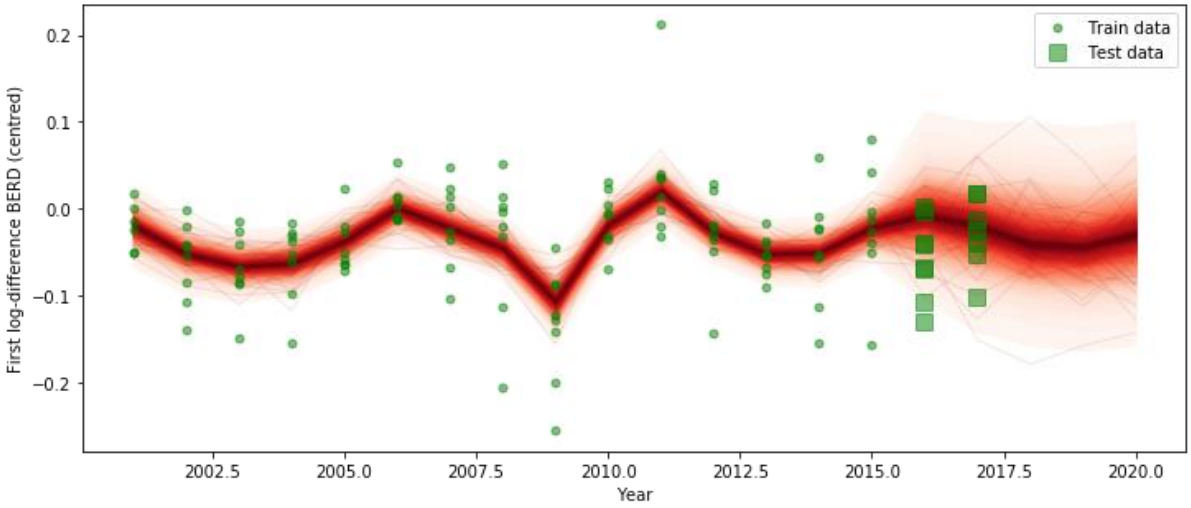


Figure 14: Predictions from the Gaussian Process model of the mean first log-difference R&D for Norway, UK, France, Germany, Belgium, Netherlands, Finland, and Sweden.

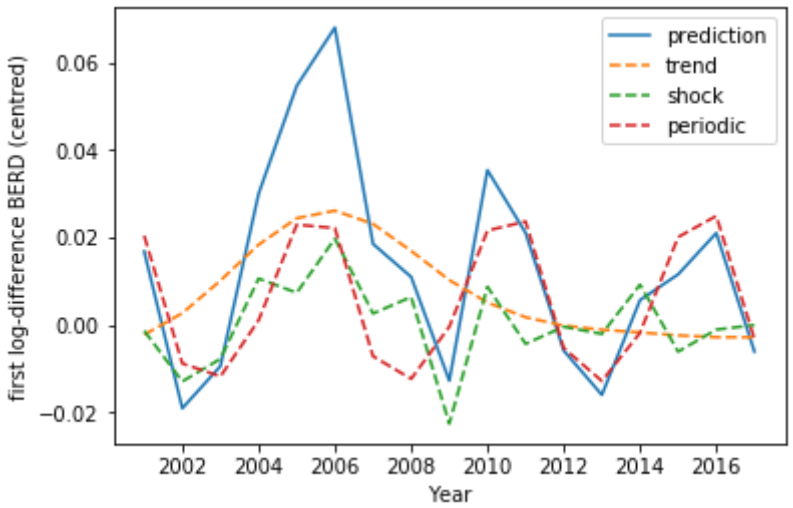


Figure 15: Prediction and trend components from the Gaussian Process model for countries entering the EU in 2004.

Figure 15 show that for countries entering the EU in 2004 (LV, LU, RO, HU, PL, ES, CZ, SI, SK) there is a large increase in R&D expenditure around entry to the EU that seems to occur as an equal combination of a long term trend, a short term shock and a peak of the periodic component - perhaps the combination of all of these is the model accounting for the fact that there is not a mid-term trend effect. It the decomposition of this time series is to be believed then the long-term trend for these countries has decreased later years suggesting there was a honeymoon period after entering the EU that has since worn off.

Future Indicators

The predictors used thus far for nowcasting have been highly traditional but in future we wish to incorporate additional variables which may include: Online job adverts (through the CEDEFOP project or through web-scraping); Economic complexity indices from the Atlas of Economic Complexity; Country import export tables (can be used to measure relatedness between countries economies to proxy how shocks in variables may propagate).

An additional reason for considering the use of Economic complexity indices such as those of Hidalgo and Hausmann (2009) are to bring in a complexity economics perspective to the nowcasting work. For example, the work of Cristelli et al. (2015) analyses the heterogeneous and non-stationary dynamics in a 2D plane of countries GDP growth and an economic complexity measure, they find that there are laminar (predictable) and chaotic (unpredictable) regimes in the plane. Incorporating both the predictive power of economic complexity indices in addition to the non-equilibrium view point of complexity science has the potential to further improve understanding, accuracy, and uncertainty estimation for nowcasting.
Figure 16 below briefly plots the economic complexity index (ECI) against our BERD and GVA variables for countries in the dataset. Our results show indications of different predictability regimes with more developed economies showing less chaotic trajectories through the phase space. The first subplot illustrates that countries are generally increasing both GVA and BERD investment, with lower GVA/BERD economies appearing to be in a more turbulent regime (though this may be because fluctuations are larger relative to the size of their economies). The second subplot indicates that as expected there appears to be some relationship between a countries' ECI and BERD in general but that this is not very smooth across time possibly due to the ECI's dependence on other countries' performance (a rank ordering may be more stable and informative) which further suggests the need to consider the relatedness of countries economies in a model.

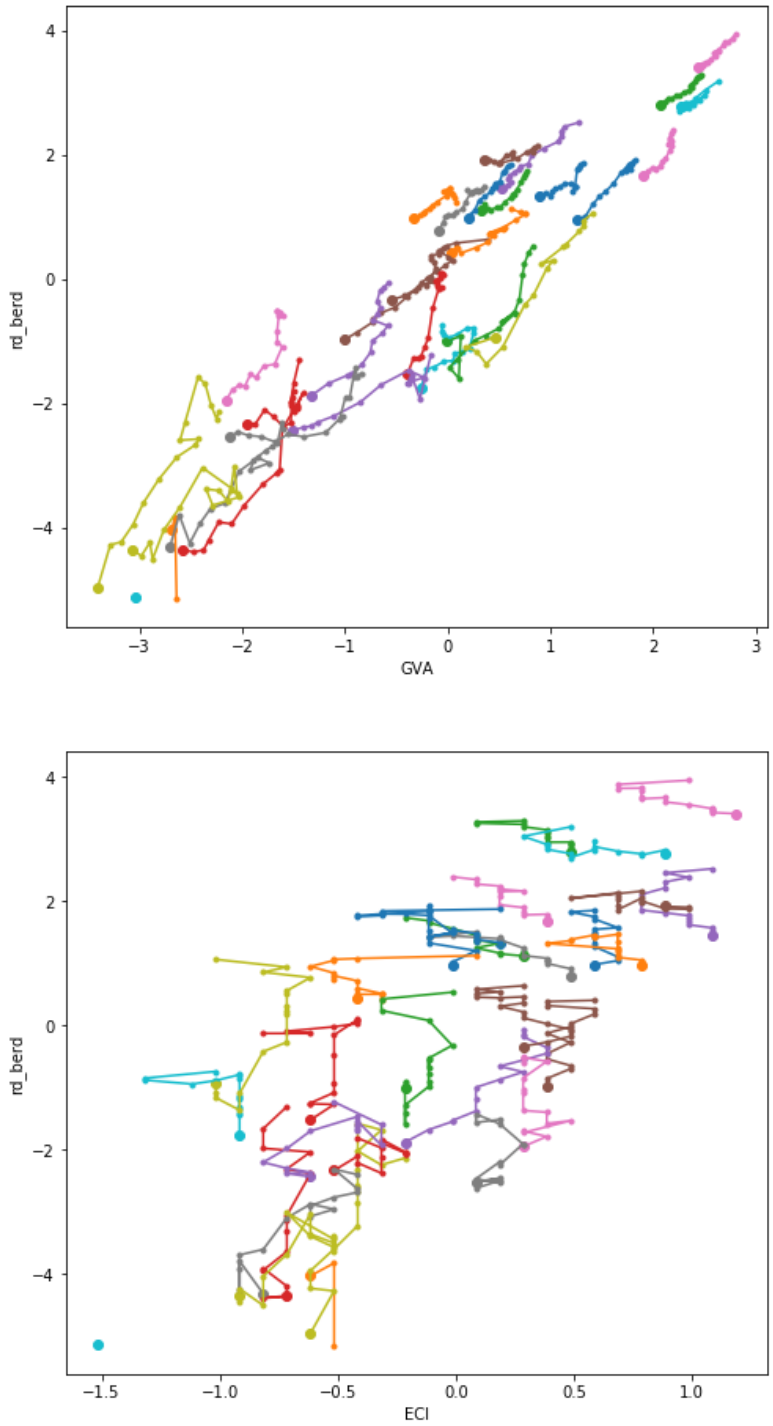


Figure 16: Countries' individual trajectories through the GVA-BERD phase space (first plot) and the ECI-BERD phase space (second plot)

Scaling up the incorporation of indicators will yield more variables than data points (the large p small n) which requires sparse models. Traditionally a LASSO (L1 penalised regression) or Elastic-Net (Combination of L1 and L2 penalised regression) are used to identify variables to select/reject by shrinking the majority of variable contributions to zero. It is widely and incorrectly known within Bayesian circles that placing Laplacian (double exponential) on a linear regression performs a Bayesian version of the LASSO. Whilst it is true that the maximum a posteriori (MAP) estimate is equivalent to that of the LASSO, the full posterior distribution does not correspond to a Bayesian

LASSO as the Laplacian distribution cannot give enough mass to both large and small coefficients simultaneously. Instead one can implement the Finnish Horseshoe model of Piironen and Vehtari (2017) which performs rigorous Bayesian sparse regression by making use of hierarchical priors that collapse the entire marginal posterior to either relevance or irrelevance.

4 Discussion and Conclusions

Bayesian framework for nowcasting shows promise for the forecasting of BERD statistics, but with more indicators and more model development required. By taking a more traditional approach to nowcasting, applying a Bayesian context, new indicators (such as startup growth, economic complexity indices, indicators derived from Meetup), and a complex systems mindset we believe there is potential to improve nowcasting accuracy for R&D statistics (and other econometric variables of interest to policymakers).

The firm level exploration of this pilot had several negative findings. Firstly, Github data do not provide reliable coverage of companies, with automated company discovery being much more challenging than the already demanding task of timely matching of firms to both patents and publications. We also observe that the large lag between R&D investment behaviour and patents/publications raises conceptual problems for their use as nowcasting indicators. In addition, matching patents and publications to companies is a highly complex and error prone process requiring access to multiple proprietary data sources (not open) and their use to apportion a firm's R&D activity to different countries is hard to verify (not trusted), and cannot provide regional granularity. A firm level approach is therefore considered to be the least promising one to carry forward in the scale-up phase as it is not open or trusted, with several other players already working on this have access to resources such as micro data that the EURITO consortium does not.

4.1 Validation and ongoing stakeholder engagement

4.2 Limitations

The models used thus far require more complexity both in the structure of the new model and the indicators used to nowcast.
The coverage analysis of scoreboard companies in section 3.1 only accounted for company subsidiaries for patents.

4.3 Considerations for scaling up

The team carrying out this pilot have the capacity to perform any scaling required, and should not require additional resources/support.

4.3.1 Complementarities with other pilots

The use of Crunchbase aligns with the inclusive-innovation pilot.
The use of papers aligns with funding analytics, though this data-source will be unlikely to be a part of a scaled up pilot.

4.3.2 Tools and data sources

The tools and data sources used do not constrain the possibility of scaling up the analysis; however more indicators are yet to be explored which could but shouldn't include closed data-sets.

4.4 Feedback obtained from third parties

5 References

Falk, M. (2006). What drives business Research and Development (R&D) intensity across Organisation for Economic Co-operation and Development (OECD) countries?. *Applied Economics*, *38*(5), 533-547.

Dernis, H. (2007). Nowcasting patent indicators.

Castle, J. L., Hendry, D. F., & Kitov, O. I. (2013). Forecasting and nowcasting macroeconomic variables: A methodological overview (No. 674). University of Oxford, Department of Economics.

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. Econometric Reviews, 26(1), 53-90.

Andreou, Elena, Eric Ghysels, and Andros Kourtellos. "Forecasting with mixed-frequency data." In The Oxford handbook of economic forecasting. 2011.

Boivin, J., & Ng, S. (2005). Understanding and comparing factor-based forecasts (No. w11285). National Bureau of Economic Research.

Mouchart, M., & Rombouts, J. V. (2005). Clustered panel data models: an efficient approach for nowcasting from poor data. International Journal of Forecasting, 21(3), 577-594.

Benages E. et al. (2018). The 2018 PREDICT Dataset Methodology. JRC Technical reports.

Isella L. (2017). https://www.conference-service.com/NTTS2017/documents/agenda/data/x_abstracts/x_abstract_28.docx

Azagra Caro, J. M., & Grablowitz, A. (2008). Data on Business R&D: comparing BERD and the Scoreboard.

Cozza, C. (2010). Measuring the internationalisation of EU corporate R&D: a novel complementary use of statistical sources. JRC Scientific and Technical Reports, Luxembourg.

Gkotsis, P., Hernandez, H., & Vezzani, A. (2016). Estimating territorial business R&D expenditures using corporate R&D and patent data (No. JRC103127). Joint Research Centre (Seville site).

Camerani, R., Rotolo, D., & Grassano, N. (2018, September). Do Firms Publish? A Cross-Sectoral Analysis of Corporate Publishing. In 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands. Centre for Science and Technology Studies (CWTS).

Kruschke, J. K. (2011). Doing Bayesian data analysis: A tutorial with R and BUGS. Burlington, MA.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. Chapman and Hall/CRC.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. Annals of internal medicine, 130(12), 995-1004.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society open science, 1(3), 140216.

Duo, Q. I. N. (1996). Bayesian econometrics: the first twenty years. Econometric Theory, 12(3), 500-516.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. PeerJ Computer Science, 2, e55.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76(1).

Becker, B. (2015). Public R&D policies and private R&D investment: A survey of the empirical evidence. Journal of Economic Surveys, 29(5), 917-942.

Doloreux, D., Shearmur, R., & Rodriguez, M. (2016). Determinants of R&D in knowledge- intensive business services firms. Economics of Innovation and New Technology, 25(4), 391- 405.

Gkotsis, P., Hernandez, H., & Vezzani, A. (2016). Estimating territorial business R&D expenditures using corporate R&D and patent data (No. JRC103127). Joint Research Centre (Seville site).

Guzman, J., & Stern, S. (2015). Nowcasting and placecasting entrepreneurial quality and performance (No. w20954). National Bureau of Economic Research.

Moat, H. S., Curme, C., Stanley, H. E., & Preis, T. (2014). Anticipating stock market movements with Google and Wikipedia. In Nonlinear phenomena in complex systems: From nano to macro scale (pp. 47-59). Springer, Dordrecht.

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. Proceedings of the national academy of sciences, 106(26), 10570-10575.

Cristelli, M., Tacchella, A., & Pietronero, L. (2015). The heterogeneous dynamics of economic complexity. PloS one, 10(2), e0117174.

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, 11(2), 5018-5051.

Daiko, T. et al. (2017). World Corporate Top R&D Investors: Industrial Property Strategies in the Digital Economy. A JRC and OECD common report. Luxembourg: Publications Office of the European Union.

Pilot 3: Technological Change Indicators

Abstract:

Quantifying the rate of technological change in a given industry or field is a prerequisite for identifying both speed and strength of technological transformations and also for supporting the management of interventions that seek to modify technological trajectories and accelerate or control the degree of technological change. A good indicator for technological change should, for example, help answering questions, such as: Are we in a period of incremental or radical change? What types of patterns of technological change exist that can help us forecast the future rate and degree of technological change? Despite the usefulness that such an indicator for technological change has, today there is no scalable cross-domain, datasource agnostic, and quantitative indicator for technological change that incorporates what is considered a crucial element of how technology changes: the combinatorial process. This pilot's focus is on R&I indicators able to capture technological change based on recombinant innovation and economic complexity approaches. Worldwide bioenergy R&D is used as the application domain to allow for an in-depth interpretation of results with domain experts. For this purpose, discussions with the EU Joint Research Centre (JRC) focusing on Energy, The Nordic Institute for Studies in Innovation, Research and Education (NIFU) have taken place, and expressions of interest from JRC Innovation and Growth have already been received.

1 Introduction

1.1 Background/context

As a driving force of technological progress, technological change has been widely conceptualised as a process of combination and recombination (Fleming and Sorenson 2001), where different new and already existing technologies are integrated resulting in a technological novelty (Strumsky et al. 2012). As such, technological change is usually manifested through the introduction of new technological functionalities into a set of existing technologies (Youn et al. 2015).

Technological change emanates from recombining and synthesizing components in a novel manner (Carnabuci and Bruggeman, 2009; Fleming, 2001) or for a new application (Henderson and Clark, 1990; Yayavaram and Ahuja, 2008). According to the economics literature, these combinations can be considered as principal sources of technology development and progress that dominate innovative activity (Youn et al., 2015). Therefore, within the combinatorial view of technological change, new and existing technologies are considered as building blocks of other new technologies (van den Oord and van Witteloostuijn, 2018; Arthur & Polak, 2006; Auerswald et al., 2000; Fleming and Sorenson, 2004).

1.1.1 Opportunity

We currently lack R&I indicators capturing technological change. Unless a technology is pre-catalogued, it is invisible to current official instruments. This means that policymakers have a blind spot when it comes to the understanding of technological change that might affect the allocation of resources, the design of regulations and the evaluation of R&I progress.

1.1.2 Application domain

The overall application domain we used in this pilot can be described as "Research and development of bioenergy solutions". In this context, bioenergy can be understood as "renewable energy made available from materials derived from biological sources". This form of energy is considered one of the key elements needed to diversify our energy production alternatives, among other things because some implementations allow producing energy with a negative carbon footprint. Despite the potential that bioenergy holds, the world is currently at a stage where we need to significantly increase the performance of current solutions, and simultaneously reduce the costs and time to market. Only in this way bioenergy technologies can have the strong environmental and economic impact that is required for large-scale implementations.

We are proposing this application domain because of:

- Its inherent potential for societal impact (market-pull).
- Worldwide availability of quality data-sources both traditional and non-traditional.
- Access to application domain experts that can help us to validate and interpret our results.

1.1.3 Flexibility of the application domain

We have designed the indicators for technological change in such a way that they can be transferred across domains with minimal need for modifications. The main challenge is that moving to other application domains might require some degree of customisation of the internal dictionaries of

taxonomies. For this reason, we have made it possible to use instead cross-domain dictionaries and taxonomies that can be obtained from platforms such as DBPedia and WikiData.

1.1.4 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

R&I indicators for technological emergence have not been “institutionalised” yet. Furthermore, most previous attempts are not based on combinatorial innovation metrics. Taking such “peripheric” indicators into the core is facilitated by the champions that we have considered for this pilot.

1.1.5 Stakeholder engagement summary

- We have engaged as stakeholders the EU Joint Research Centre (JRC) (Energy and transport area), The Nordic Institute for Studies in Innovation, Research and Education (NIFU), the Research Policy journal through his editor John Walsh and assistant professor Stanko Skec from the University of Zagreb whom we worked with on the fundamentals of technological change indicators.

- The response we have obtained so far from the interactions with the stakeholders indicates that 1) the approach we are taking is not something they have seen previously applied in this context 2) they believe the database coverage is appropriate and sufficiently representative of the application domain and 3) the results align with their understanding of the overall patterns of technological change in the sector.

- Feedback highlights from the knowledge stakeholder workshop (KSW) include a) the possibility of exploring different technological change patterns per country in order to analyse stages of development, and/or technological specialisation at the regional or organizational levels, b) we should evaluate means to go from our originally descriptive to a potentially predictive analysis, taking advantage of the observed trends, and c) we could analyse how technological changes exhibit different patterns depending on the data source used. For example, compare the technological change pattern found within the corpus of patents with the patterns of scientific publications and/or grants.

1.2 Relevance to RITO criteria

1.2.1 Relevant

The pilot tackles an issue that has been already mentioned as something that is necessary to improve in the context of R&D indicators, and it is also of high societal relevance.

1.2.2 Inclusive

We are setting a worldwide scope, including developing and developed countries in the data-pool. As our emphasis is also in network structure, not scale, we believe this helps to highlight unique/innovative solutions that emerge from countries that produce a relatively small volume of outputs. Furthermore, we have combined industrial and academic outputs and the data inputs can be expanded to include data related to SMEs and any other data-source available.

1.2.3 Timely

Our addition of data sources other than patents and publications allow us to identify signals that might be expressed earlier on in other sources. Furthermore, focusing not just on keywords, but on the combination of technological building blocks allows earlier identification of unique technological changes.

1.2.4 Trusted

Combining traditional/curated data sources with non-traditional sources allows us to leverage the trust put in official sources whilst also enjoying the complementary advantages of alternative sources.

1.2.5 Open

All our proposed methods and data for this proposed pilot are already publicly available (at least in some form). We will also document the analytical steps taken so that they are reproducible.

1.3 Research/policy questions

- What is the year-to-year rate of technological change?
- For example, are we in a relatively stable and incremental period or in a period where the rates of change seem to point to radical departures from our previous technological trajectories?
- Which technologies are the biggest contributors to the overall pattern of technological change measured in a given period?
- How unique/new is a given technology?

2 Methodology

2.1 Data sources

In the context of this pilot we have used the following data sources:

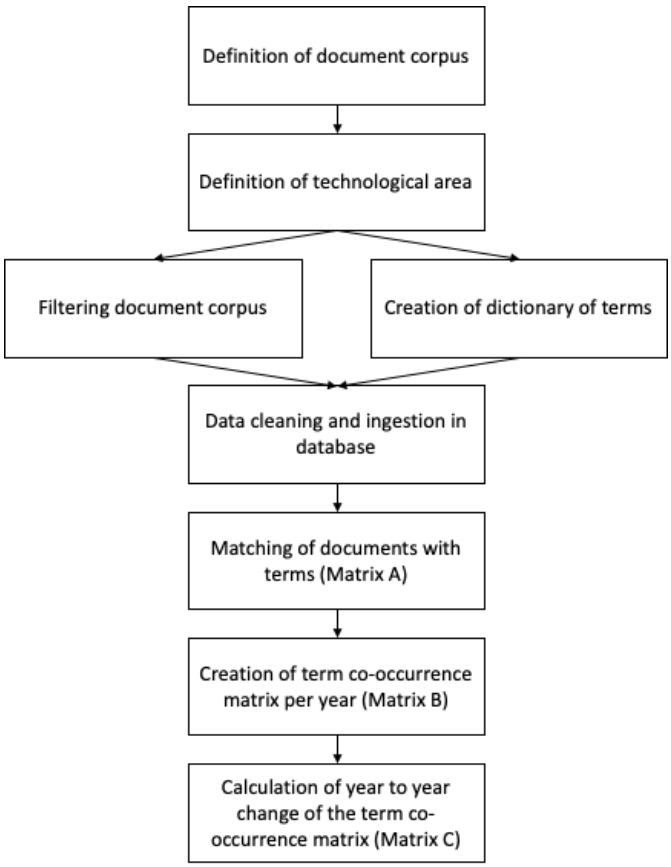
Source	Status	Observations
Scientific Publications	We have analysed worldwide bioenergy and biofuel publications from Web of Science.	Web of Science requires a subscription (that most universities have) and allows data mining for research purposes. However for other uses or situations where a license is not available we can substitute Web of Science with open access publication repositories.
Research funding	We have analysed H2020 funding data	It is possible to include data from other EU funding organisations
Data generated by previous EU projects working on bioenergy and worldwide bioenergy agencies	We have included data extracted from Bioenergy2020+, ETIP Bioenergy, Biofuel Digest and Genscape databases. These data sources exist as openly accessible websites that contain repositories with bioenergy projects, feedstocks and processing technologies.	The heterogeneous set of data-sources used in this category might need to be streamlined before scaling up
Patents	We have analysed worldwide bioenergy and biofuel publications extracted from Derwent.	Derwent requires a subscription (which some universities have) and allows data mining for research purposes. However for other uses or situations where a license is not available we can substitute Derwent with open patent publication databases (e.g. patent lens).

2.2 Method

Consistent with a combinatorial perspective on technological change, the proposed method uses the occurrence and co-occurrence of terms within a corpus of documents (Feldman and Sanger 2007) to describe different combinatorial configurations within the document corpus of R&D-related records. In our approach terms are text strings of one or more words, also called n-grams (Dale, Moisl and Somers

2000), that represent technology-relevant entities such as production inputs (e.g. barley straw), processing technologies (e.g. pyrolysis) and outputs (e.g. biogas). More specifically, we use changes over time in the occurrences and co-occurrences of such terms as a proxy for technological changes. In this approach, the occurrences and co-occurrences of selected terms within documents are used to build adjacency matrices that store the weighted combinations of those terms and term-pairs for time period. Such matrices serve 1) as a description of the combinations of terms that have been explored in a given period of time and 2) to calculate configurational changes in the matrix from one period to the next. An overview of the key steps in the process is provided in figure 1 below.

Figure 1: Overview of the developed method



2.3 Documentation

The code developed and employed in this pilot is available in the form of a Jupiter which also provides the means to generate most of the visualisations presented in the results section.

The data used in this pilot is available in a Github repository that contains both the terms and the full set of document data employed.

Finally, a set of slides with the method and results of the pilot were also produced for dissemination.

The code, slides and data are available at:

https://github.com/EURITO/wp2pilots/tree/master/Pilot3

3 Results

3.1 Outputs - Technological changes in bioenergy R&D

In order to demonstrate our method, we applied it to measure technological changes in the context of the research and development work in the area of bioenergy solutions. Research and development of bioenergy solutions is an active technological field, with R&D-related records starting from the mid-seventies (GUPTA VK, TUOHY MG, KUBICEK CP, SADDLER J, XU F, 2014; Web of Science) and that has experienced a significant increase in the volume of R&D documents in the last two decades (see fig 3, logarithmic scale).

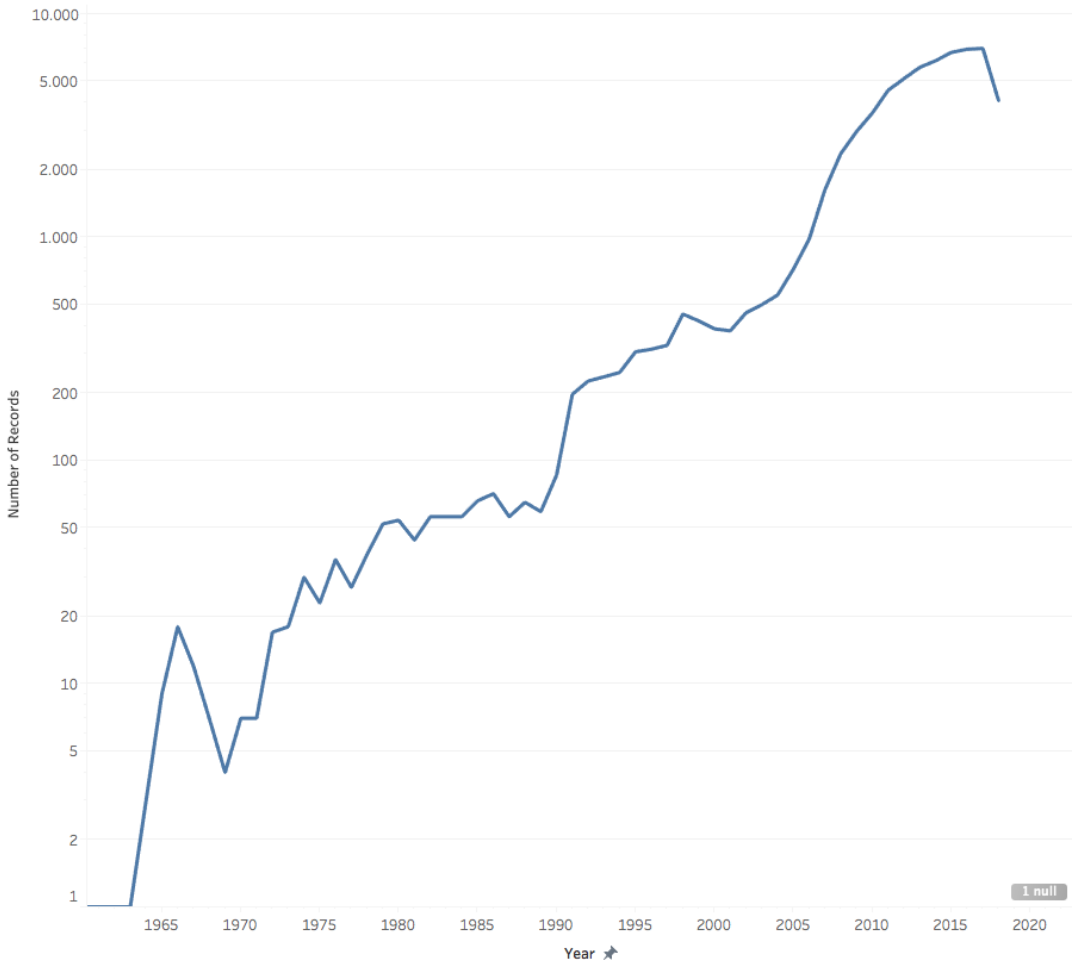


FIG 3: Volume of R&D-related documents within the bioenergy field (logarithmic scale)

Bioenergy R&D is concerned with the generation of renewable energy using materials derived from biological sources such as biomass. One of the main areas within this technological space is the production of biofuels. During the lifetime of this field, there have been several technological changes related to aspects such as the production inputs used (the feedstocks) and the processing technologies utilised). The main drivers for these changes are connected to the social, environmental and economic sustainability of biofuels. These drivers have translated into pressures to increase the speed and volume of the production of greener and cheaper biofuels that can become viable alternatives to fossil fuels. Retrospectively, such changes have been characterised in terms of what is now known as four biofuel generations (Aro 2016).

DATA SOURCES AND CREATION OF A DOCUMENT CORPUS

Following our method, to achieve a broad coverage of relevant research and development document sources, we have text-mined a diverse set of historical records of technology-related R&D activity. This records include patents, scientific publications, official EU project descriptions (Cordis database), as well as databases that include Bioenergy2020+, ETIP Bioenergy, Biofuel Digest and Genscape which contain descriptions about biofuel-specific projects and worldwide biofuel facilities.

To focus on those results that are most clearly associated to bioenergy or biofuels, we have used the following text string as filter for all the data-sources: [“biofuel* OR bio-fuel* OR "bio fuel*" OR bioenerg* OR "bio energ*" OR bio-energ*”]. This text string was selected based on the amount and quality of the results it provided, which after manual examination showed a good balance between recall and precision. After applying the filter, the number of documents that we include in our analysis per data source is:

- Scientific publications (Web of Science): 58.239 documents
- Patents (Derwent Innovations Index): 6.570 documents
- Official EU project descriptions (Cordis database): 692 documents
- Biofuel facilities and projects (Bioenergy2020+, ETIP Bioenergy, Biofuel Digest and Genscape databases): 1.647 documents

Although the majority of the records come from scientific publications, the additional coverage that patents, projects, and biofuel facilities provide, allow us to capture terms and term-pairs that are widely used outside academic circles but, relatively less represented in scientific publications. In addition, since the year-to-year analysis is not affected by absolute volume, but rather by the overall configuration of the term matrix within a yearly variations in volume between document sources are less problematic.

3.2 Findings

Interpretation of the indicator of technological changes applied to the bioenergy R&D case

The results shown in the matrix that stores the RV-coefficients for all year pairs allow us to identify four main findings:

1) As intuition would suggest, in general, when the time between two years increases their similarity decreases. This is an indication that our method is able to capture the theoretically expected macro-behaviour of technological change, which predicts that over time the accumulation of year-to-year changes (both incremental and radical) should lead to an increase in the accumulation of technological changes over time (Parayil 1993). For example, an examination of the RV-coefficient curve for the year 2017 against all other years, see figure 2 below, shows that with few exceptions the farther we move from 2017 the lower is the RV coefficient.

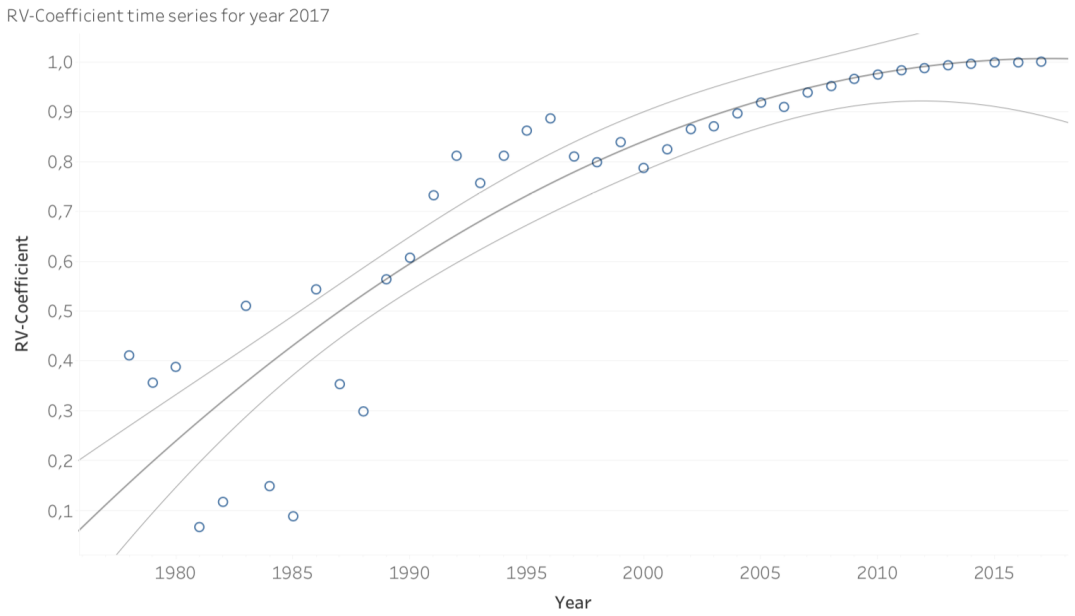


Figure 2: RV-Coefficient time series for year 2017

2) As shown in figure 3 below, year-to-year similarity measures are relatively low from one year to the next in earlier periods and are higher in later periods. This indicates that in earlier periods, i.e. 1978-1990, year-to-year configurational changes in the matrix that stores the combinatorial possibilities are of larger magnitude and more frequent. As the time passes, i.e. 1990 onwards, year-to-year changes become smaller, which can be interpreted as a sign that overall bioenergy R&D is settling into more stable technological configurations.

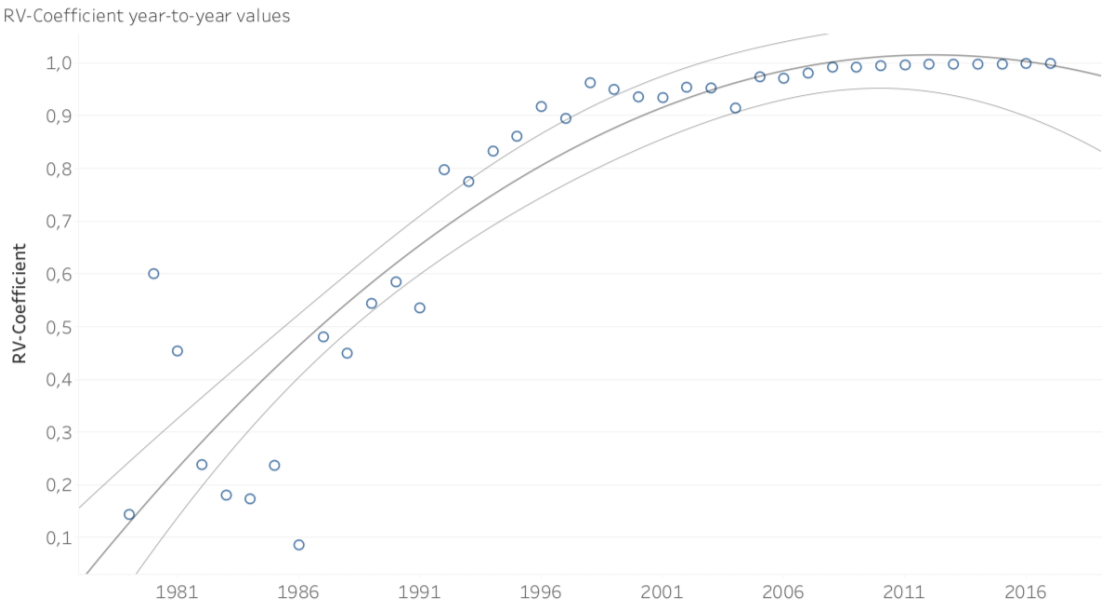


Figure 3: Year-to-year similarity

3) Year-to-year similarity measures are generally much lower and more rapidly changing in earlier years than in later years. We see this as a marked difference in RV Coefficients in the period pre 1991 (average RV-Coefficient of 0,31) and post year 1991 (average RV-coefficient of 0,88). One element that influences this behaviour is the lower volume of documents in earlier years, which means that the chances available to explore all potential combinatorial possibilities are reduced. However, this is not an artifact of the sampling method or the data-sources available, instead it shows that as the volume of R&D activity increases it is possible to explore a wider set of combinatorial possibilities, which in turns translate in more stable year-to-year similarity measures. A second element that influences this behaviour is the more experimental and uncertain nature of early explorations of a new technological field. This means that shifts in early periods will translate into larger configurational changes when compared to later periods. One driver for this is the relatively small volume of overall R&D activity earlier on, which makes each change a more significant percentage of total R&D activity. In contrast, in later periods the larger volume of R&D activity can be distributed into multiple parallel areas of research.

4) It is possible to identify blocks/groups of years (clusters) with higher and more stable year-to-year similarity that are interrupted by changes that break the high similarity found within the block and after such change a new block emerges. Using a hierarchical clustering analysis of matrix C, shown in figure 4, we can identify the following tree structure describing the year intervals that have high similarity within interval and low similarity outside of the interval:

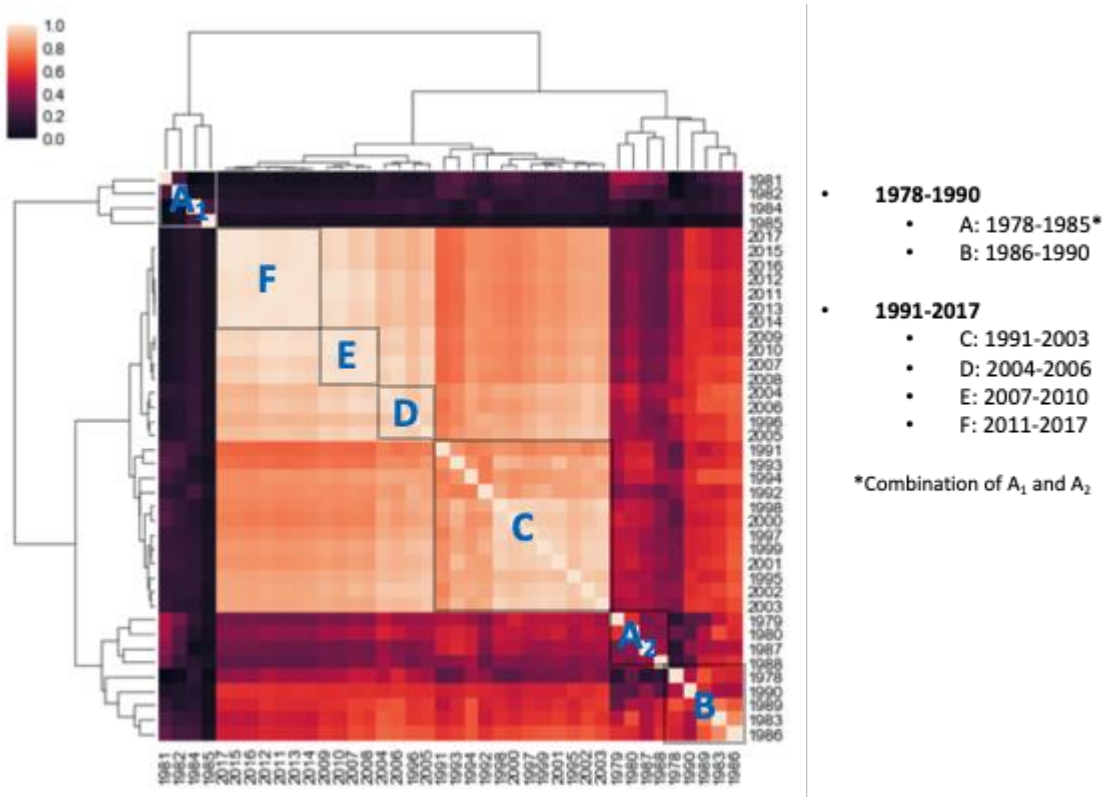


Figure 4: Hierarchical clustering analysis of years

4 Discussion and Conclusions

The proposed methodology and indicator allows to quantify technological change taking into account a combinatorial perspective. In contrast, previous approaches aimed at quantifying technological change often rely on measurements of the amount of terms in documents, or predefined categories. Although there have been a number of measures and metrics to quantify technological change that are "objective and reproducible" (e.g. review by Suominen 2013), such approaches are often tailored according to the needs of specific application domains without providing general guidelines for applying them to different technological fields (e.g. Goldfarb, 2005). Therefore, although previous approaches offer some advantages in exploring technological change, their drawbacks may lead to a potentially biased and distorted view of technology dynamics and evolution (e.g. Funk and Smith, 2016). For example, an increase in R&D funding might result in a higher number of publications and patents without the expected technological diversification (Kook et al. 2017).

In comparison to previous approaches for the analysis of technological changes, the proposed network-based methodology allows further investigation of the combinatorial possibilities on a macro level and can serve as a suitable platform for complementary analyses of alternative technological change measurement approaches. The notion of network-based representation and understanding of technological change also provides the opportunity to explore the evolution of preexisting configurations and subsequent additions, replacements, etc.

4.1 Validation and ongoing stakeholder engagement

- We will continue our engagement with The Nordic Institute for Studies in Innovation, Research and Education (NIFU) and the rest of the stakeholders.
- In order to validate the empirical results of our method we will involve bioenergy and biofuel researchers at DTU that can compare the patterns that we found against their own experience and the literature base.
- Our plan is that the rest of the activities aimed at validation and ongoing stakeholder engagement will be primarily carried on through bilateral discussions that will allow us to go into the high-level of detail needed at this stage of the work.

4.2 General limitations

- The developed indicator does not capture impact, performance, cost, etc. Focus is on early stage R&D and on overall technological change
- The interpretation of terms co-occurrence is valid only on aggregate. This means that the observation of a change in a pair of terms such as "Algae - Biodiesel" should not be interpreted in isolation. To correctly interpret changes in the usages of terms or term pairs is necessary to consider how all other terms and/or term pairs change and the overall volume of documents. At the aggregate level, the indicator for technological change that we developed takes into consideration both the volume of documents and entire set of changes a whole.
- The relation between volume of document records and opportunity to explore space of combinatorial possibilities requires additional work to avoid misinterpretations

4.3 Considerations for scaling up

The main challenge to scaling up relates to the creation of standard cross-domain taxonomies or dictionaries. The method has been tested with different combinations of dictionaries and is robust to changes but all the tests have been performed within the context of bioenergy R&D. If we decide to move elements of this pilot to the scaled-up cases the first task will be to build large cross-domain dictionaries extracting entities from sources such as DBpedia and WikiData.

4.3.1 Complementarities with other pilots

- This pilot is highly complementary with the efforts carried out in pilot #1, emerging technology ecosystems in AI, as it allows bridging the micro-level patterns identified in pilot 1 against the macro-level technological changes identified in this pilot.

- There is an interesting opportunity to combine the usage of standards promoted in pilot "Pilot 4 Standards as indicators for the diffusion of innovation". This can be done by integrating a corpus of standards and/or regulations as an additional datasource within the same methodology and running the same analyses.

4.3.2 Tools and data sources

The main tools used in this pilot are a Graph Database running on on Neo4j, Python and Jupyter Notebooks (where we have documented each step of the process).

References

Abercrombie, R.K., Udoeyop, A.W., Schlicher, B.G., 2012. A study of scientometric methods to identify emerging technologies via modeling of milestones. Scientometrics 91, 327–342.

Adegbesan, J.A., Ricart, J.E., 2007. What Do We Really Know about When Technological Innovation Improves Performance (and When it Does Not)? SSRN Electron. J. 3.

Lee, J., Berente, N., 2013. The era of incremental change in the technology innovation life cycle: An analysis of the automotive emission control industry. Res. Policy 42, 1469–1481.

van den Oord, A.., 2010. The Ecology of Technology: The Co-Evolution of Technology and Organization.

Kook, S.H., Kim, K.H., Lee, C., 2017. Dynamic technological diversification and its impact on firms' performance An empirical analysis of Korean IT firms. Sustain. 9.

Funk, R.J., Owen-Smith, J., 2017. A Dynamic Network Measure of Technological Change. Manage. Sci. 63, 791–817.

Goldfarb, B., 2005. Diffusion of general-purpose technologies: Understanding patterns in the electrification of US Manufacturing 1880-1930. Ind. Corp. Chang. 14, 745–773.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strateg. Manag. J. 25, 909–928.

Auerswald, P., Kauffman, S.A., Lobo, J., Shell, K., 2000. The production recipes approach to modeling technological innovation: An application to learning by doing. J. Econ. Dyn. Control 24, 389–450.

Arthur, W.B., Polak, W., 2006. The evolution of technology within a simple computer model. Complexity 11, 23–31.

Yayavaram, S., Ahuja, G., 2008. Decomposability in Knowledge Structures and Its Impact on the Usefulness of Inventions and Knowledge-base Malleability. Adm. Sci. Q. 53, 333–362.

Carnabuci, G., Bruggeman, J., 2009. Knowledge Specialization, Knowledge Brokerage and the Uneven Growth of Technology Domains. Soc. Forces 88, 607–641.

Fleming, L., Sorenson, O., 2001. Technology as a complex adaptive system: Evidence from patent data. Res. Policy 30, 1019–1039.

Arthur, W.B., 2009. The nature of technology: what it is and how it evolves. Free Press.

Aro, E.M., 2016. From first generation biofuels to advanced solar biofuels. Ambio 45, 24–31.

Nadeau, D., 2007. A survey of named entity recognition and classification. Linguist. Investig. 3–26.

Gupta, V.K., Tuohy, M.G., Kubicek, C.P., Saddler, J., Xu, F. (Eds.), 2014. Bioenergy Research: Advances and Applications. Elsevier.

van den Oord, A., van Witteloostuijn, A., 2018. A multi-level model of emerging technology: An empirical study of the evolution of biotechnology from 1976 to 2003. PLoS One 13, e0197024.

Ramsay, J.O., ten Berge, J., Styan, G.P.H., 1984. Matrix correlation. Psychometrika 49, 403–423.

Everett, M.G., Borgatti, S.P., 2013. The dual-projection approach for two-mode networks. Soc. Networks 35, 204–210.

Josse, J., Pagès, J., Husson, F., 2008. Testing the significance of the RV coefficient. Comput. Stat. Data Anal. 53, 82–91.

Robert, P., Escoufier, Y., 1976. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. Appl. Stat. 25, 257.

Youn, H., Strumsky, D., Bettencourt, L.M.A., Lobo, J., 2015. Invention as a combinatorial process: evidence from US patents. J. R. Soc. Interface 12, 20150272–20150272.

Strumsky, D., Lobo, J., van der Leeuw, S., 2012. Using patent technology codes to study technological change. Econ. Innov. New Technol. 21, 267–286.

Smilde, A.K., Kiers, H.A.L., Bijlsma, S., Rubingh, C.M., Van Erk, M.J., 2009. Matrix correlations for high-dimensional data: The modified RV-coefficient. Bioinformatics 25, 401–405.

Abdi, H., 2007. RV Coefficient and Congruence Coefficient. Encycl. Meas. Stat. 849–853.

Guthrie, S., Wamae, W., Diepeveen, S., Wooding, S., Grant, J., Europe, R., 2013. Measuring Research: A Guide to Research Evaluation Frameworks and Tools, RAND Monographs.

National Research Council, 2014. Capturing Change in Science, Technology, and Innovation. National Academies Press, Washington, D.C.

Buckland, M., Gey, F., 1994. The relationship between Recall and Precision. J. Am. Soc. Inf. Sci. 45, 12–19.

O'Keeffe, A., McCarthy, M., 2012. The Routledge handbook of corpus linguistics. Routledge, Milton Park Abingdon Oxon ;;New York.

Feldman, R., Sanger, J., 2007. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, New York, NY.

Wachsmuth, H., 2015. Text analysis pipelines: towards ad-hoc large scale text mining.

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.

Henderson, R.M., Clark, K.B., 1990. Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. Adm. Sci. Q. 35, 9.

Pilot 4: Standards as Innovation Diffusion Indicators

Abstract
The pilot on standards has the objective to test the usability of certifications of selected international management standards, like on quality and environmental management, released by the International Organization for Standardization (ISO) as indicators for the diffusion of innovation. Descriptive statistics of time series and distributions among sectors are displayed and regional heatmaps of indicators are constructed as certifications per number of companies in the EU28. The data collected and provided by ISO have been validated in the case of Germany with data disclosed at the homepages of 1.2 million German companies out of the around 3 million German companies. Considerations for scaling up are presented.

1. Introduction

1.1 Background/context

1.1.1 Theoretical Foundation

Based on Schumpeter (1939), who introduced the distinction between invention and innovation as at the market successfully deployed invention, we assume that standards are indeed innovations, because standards are only developed, if there is a sufficient demand at the market, i.e. an immediate potential for their diffusion. Furthermore, the interested stakeholders have to invest significant resources, which can be up to ten million Euros in case of more complex standards (Spring and Weiss 1995). Consequently, the theory of diffusion of innovation (e.g. according to Rogers 1995 or Geroski 2000) can be applied on standards (Tuczek et al. 2018) and eventually also put in the context of macroeconomic growth models (Acemoglu et al. 2012). Standards are used for the differentiation in competition with other companies for the signaling (Spence 1973) of qualitatively superior or environmentally beneficial processes and products (Rysman et al. 2018). Compatibility standards are not isolated, but closely connected innovations, which requires to consider network effects (Weitzel et al. 2006). On the one hand, these network effects foster the diffusion of this category of standards, on the other hand sufficient critical masses of users are necessary for initiating diffusion (Arthur 1989). Therefore, we can rely on the insights of social network analysis for the investigation of the diffusion of standards (Guler et al. 2002).

The application of the diffusion theory has to be complemented by another approach. Nelson and Winter (1982) have introduced routines as an important element in their evolutionary model for explaining economic change. In particular, management standards represent routines, which may have an influence on the innovation strategies and activities of companies. This type of standards represents primarily process innovations, which according to Utterback and Abernathy (1975) have an important role in the later stages of technology cycles to achieve efficiency gains. These process innovations could also build the basis for a new generation of product innovations. Therefore, standards can also be used as indicators for identifying specific phases of technology cycles (Swann 2000), e. g. as expression of the establishment of dominant designs (Suarez 2004). The signaling function of standards and certificates in particular of quality (Terlaak and King 2006) and other characteristics, which are difficult to perceive, completes their functions for establishing innovations at the market, but also for increasing of market efficiency (Akerlof 1970). Although the relevance of standards for platforms is stressed (e. g. Gawer 2014), there exists still a research gap. In addition to these company and technology focused approaches, standards play also an important role for the implementation of systemic innovations. They can promote niche technologies in socio-technological systems (Geels 2002) to their breakthrough by the inclusion

all relevant stakeholders in their development. In the European Union, it has to be considered, that standards are the results of the self-regulation of the economy and are sometimes used for the specification of legal regulations. Therefore, they are an important element of the regulatory framework conditions and technological infrastructure (Tassey 2000). Finally, the trend from the originally technology-focused standardisation towards services and eventually the various dimensions of sustainability have to be highlighted. Therefore, standards cover not only non-technical, but also social innovatio

1.2 Opportunity

The majority of research and innovation indicators focus on the creation of inventions and innovations. Only a few output and even fewer indicators measure the diffusion of innovations. However, only the diffusion of innovation generates real economic impact. As explained above, standards contribute to the diffusion of innovation (Robertson & Gatignon 1986) and have therefore obviously an impact, e.g. on economic growth, trade, shares in values chains, but also profits (e.g. Wakke et al. 2016), safety (Lim and Prakash 2017), and the environment (Prakash and Potoski 2014). Consequently, standards in general, but also standardisation have the potential to be used as innovation indicators (see a recent survey in Blind 2019a).

1.2.1 Application domain

In contrast to other pilots, which focus on specific scientific or technological application domains, like "Artificial Intelligence" or "Research and development of bioenergy solutions", we concentrate on international management standards and their certifications across economies and sectors. They have mainly been used to explain international trade flows (Clougherty and Grajek 2008, 2014; Blind et al. 2018), but also the diffusion of environmental innovations (Lim and Prakash 2014). However, the certifications of different types of standards implemented in companies can be considered as important types of routines defined as regular and predictable firm patterns (Nelson and Winter 1982) and therefore as process innovations (see for more conceptual elaborations Blind 2019b). At the macro level, standards are important technology infrastructure (Tassey 2000) and contribute to innovation-friendly framework conditions (Blind et al. 2017).

In this pilot, we explore how certifications of international management standards can be used to track process innovations to improve productivity and quality, but also addressing sustainability issues in the EU.

We are proposing the certifications of international management standards, because:

- They are indicators for the diffusion of process innovations.
- They are of high quality and available not only for the Member States of the EU, but for more than 150 countries at the national, the sectoral level and over time.
- We have access to application domain institutions and experts that can help us to validate and interpret our results.

1.2.2 Flexibility of the application domain

Due to the cross-sector and cross-country character of international management standards, there is indeed less consideration related to the flexibility of the application domain. However, in contrast to other types of documents, like scientific publications and patents, there is less flexibility, but also less connectedness to other types of documents, due to the in general missing information of authorship.

1.2.3 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

Standards have only been recently "institutionalised" yet, like by their integration of the OECD into its Oslo Manual (OECD/EUROSTAT 2018). Taking standards as originally "peripheric" indicators into the core has been facilitated by the OECD and the EC as champions that we have considered for this pilot.

1.2.4 Stakeholder engagement summary

The OECD has realised the need to expand their definition of innovation and has – also based on the input provided by us – consequently included standardisation in the 4th edition of the Oslo Manual (OECD/EUROSTAT 2018). In detail, it claims the adoption of Total Quality Management (TQM), part of the ISO 9001 family of standards is another organisational capability that is closely related to innovation. Furthermore, standards have been included in the sources of ideas and information for innovation and the elements of the external environment for business innovation. Compliance with standards is also within the innovation objectives as well as the innovation drivers. In addition, the active contribution to standards is also perceived as innovation objective. Overall, standards and standardisation play an important coordination role in many markets and can influence the characteristics of product and business process innovations.

In parallel, the European Commission includes standards as an output indicator in the Interim Evaluation of Horizon 2020. In addition, there is a strong interest in DG GROW in the growth and trade related impacts of standards and in DG CONNECT in the important role of standards for the interoperability of ICT.

Finally, the team is in contact with national, like in Germany DIN, European, like CEN, and international standardisation bodies, like ISO, to discuss and promote standards as innovation indicators. In particular, ISO responded to one of our requests and made sector-specific information in their countries available on their homepage.

1.3 Relevance to RITO criteria

1.3.1 Relevant

Policymakers need detailed and timely information about the diffusion of standards as indicators both of the self-regulatory infrastructure and of process innovations in order to support them and evaluate the impact of self-regulatory interventions. The pilot provides this information for quality, environmental, IT security and energy efficiency management. In particular, the latter two are emerging relevant topics.

1.3.2 Inclusive

International standards released by ISO are developed in a consensus process involving all relevant stakeholders. In addition, the data provided by ISO covers more than 150 countries worldwide. In addition, the international management standards address not only the traditional quality management, but also topics, like environmental protection, energy efficiency or IT security, which can be also summarized under inclusive.

1.3.3 Timely

The certification data have a time lag of one year, i.e. published in 2018 they cover the certifications issued in 2017.

1.3.4 Trusted

The data is provided by ISO, an UN organisation. In addition, standardisation processes follow a strict consensual process. Findings are validated and triangulated with data collected from companies' homepages.

1.3.5 Open

The pilot uses in general open data and its findings are transparent and reproducible.

1.4 Research/policy questions

The pilot addresses the question of whether standards can be used as an indicator for the diffusion of innovation. Following the positive outcome of a feasibility test, it is possible to address the following questions:

- What is the diffusion rate of (management) standards over time?
- What is the diffusion rate of (management) standards across countries?
- How do the indicators correlate with the indicators of the European Innovation Scoreboard?

2. Methodology

2.1 Definitions

Before we describe the data sources and the methodology, we have to define some key terms:

- **Standardisation -** a activity of establishing, with regard to actual or potential problems, provision for common and repeated use, aimed at the achievement of the optimum degree of order in a given context (EN 45020:2006)
- **Standard** - a document established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context (EN 45020:2006)
- **Management system standards** (MSS) – guidelines helping organizations to improve their performance by specifying repeatable steps that organizations consciously implement to achieve their goals and objectives, and to create an organizational culture that reflexively engages in a continuous cycle of self-evaluation, correction and improvement of operations and processes through heightened employee awareness and management leadership and commitment
- **Certification** – the provision by an independent body of written assurance (a certificate) that the product, service or system in question meets specific requirements.

Table 1: Most relevant management standards

ISO 9001	Quality Management to ensure that products and services meet customers' needs
ISO 14001	Environmental Management to improve your environmental performance
ISO 50001	Energy Management to help organizations manage their energy performance
ISO/IEC 27001	Information security management to ensure that organizations' information is secure
ISO 22000	Food safety management: Inspire confidence in your food products
ISO 20121	Sustainable events to manage the social, economic and environmental impacts of events
ISO 13485	Medical devices
ISO 22301	Societal Security - Business Continuity Management Systems
ISO 20000	Information Technology - Service Management
ISO 28000	Specification for security management systems for the supply chain
ISO 39001	Road traffic safety (RTS) management systems

2.2 Data sources

Although the metadata, but no information about the diffusion, of around 20.000 ISO standards is available, we focus in the pilot on the data about the certifications based on the most important ISO

management standards, like the more than one million certifications based on the ISO 9001 standard on quality management and the more than 300,000 certifications related to ISO 14000, the environmental management standard. The application domain is therefore focused on the management standards released by ISO. Therefore, the application domain is very generic with the restriction that ISO standards do not consider electrotechnical topics in the narrower sense. In order to normalize the data per country, we use the data about the number of companies in the EU Member States provided by EUROSTAT in order to calculate simple intensities. We analysed the data primarily in a descriptive way in order to show the usability of standards and related certifications as a diffusion indicator, but we will also produce heat maps by countries. However, we put the new indicator also in the context of the indicators used for the European Innovation Scoreboard (EC 2018).

Every year ISO performs a survey of certifications to its most popular management system standards among its more than 150 member states. The surveys reveal the number of valid certificates to ISO management standards, e.g. as ISO 9001 and ISO 14001, reported for each country, each year and each sector. ISO itself does not provide certification services. Organizations, i.e. mostly companies, interested to get certified based an ISO standard must contact an independent certification body. The ISO Survey counts the number of certificates issued by certification bodies, that have been accredited by members of the International Accreditation Forum (IAF). This means that these certification bodies must follow themselves a set of certifications.

The full survey data is available in Excel format at the following homepage https://www.iso.org/the-iso-survey.html

Despite high efforts to display consistent results, there are fluctuations – in addition to the variability in numbers of certificates reported each year by individual certification bodies reflecting the diffusion of the various management standards – in the number of certificates from year to year due to several reasons. First, the participation of certification bodies at the national level is not always consistent, i.e. some certification bodies contribute to the survey in one year but not in the next and there are also changes in the reporting units, e.g. company level vs. firm sites. Secondly, there is always the chance that new certification bodies are created, which are then also eventually invited to report the number of the certifications, they issue. However, meanwhile the quality of the reported data has been consolidated on a high level, which lead to use the data both for scientific publications and to inform policy makers.

Table 2: Worldwide number of most relevant international management standards (Source: ISO 2018)

Standard	**Number of certificates in 2017**
ISO 9001	1,058,504
ISO 14001	362,610
ISO 50001	22,870
ISO 27001	39,501
ISO 22000	32,722
ISO 13485	31,520
ISO 22301	4,281
ISO 20000-1	5,005
ISO 28000	494
ISO 39001	620
TOTAL	**1,558,127**

2.3 Documentation

The data used in this pilot will be made available in the Github repository https://github.com/EURITO/wp2pilots that contains both the terms and the full set of document data employed.

3. Results

The results of our analyses will be presented according to the three dimensions time, sector and country.

3.1 Time trends in EU28

The number of ISO 9001 certifications related to quality management have started to stagnate and even decrease from over 400.000 in 2010 to around 350.000 in 2017 (Figure 1). It has to be noted that the first version of the standard has been released in 1987, more than 30 years ago. At the global level, the decrease is rather marginal with -4% and an overall number of above one million certificates. This reduced relevance among companies and other organisations is also reflected by the shrinking number of scientific papers focusing on ISO 9001 (Pohle et al. 2018). Despite the fact that the environmental management standard ISO 14001 has been released almost ten years later, the number of ISO 14001 certifications is also stagnating at a level just above 100.000 in the 28 Member States of the EU. This sluggishness is in contrast to the continuing moderate growth of 5% at the global level and the slightly, but still increasing number of scientific publications addressing this environmental management standard. Overall, these two standards have meanwhile obviously reached a saturated level in their diffusion after their take-offs in the 1990ies in case of ISO 9001 respectively the first decade of the 21st century in case of ISO 14001.

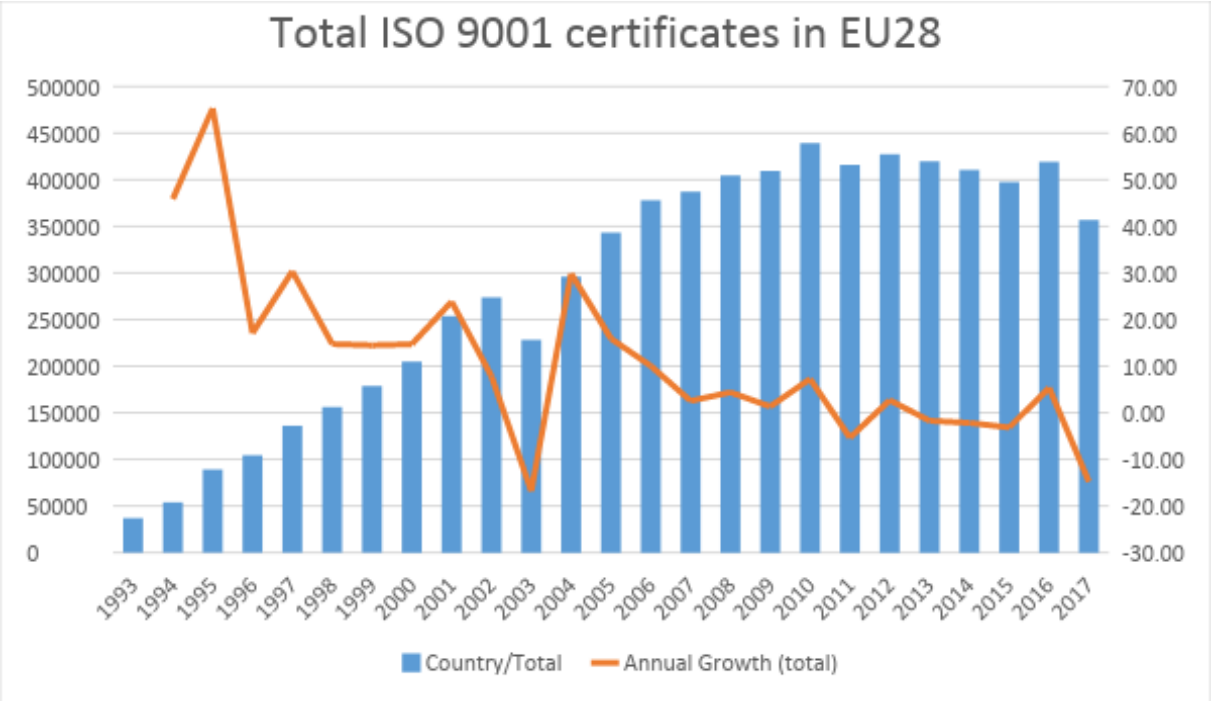


FIG 1: Number of ISO 9001 certificates in EU28

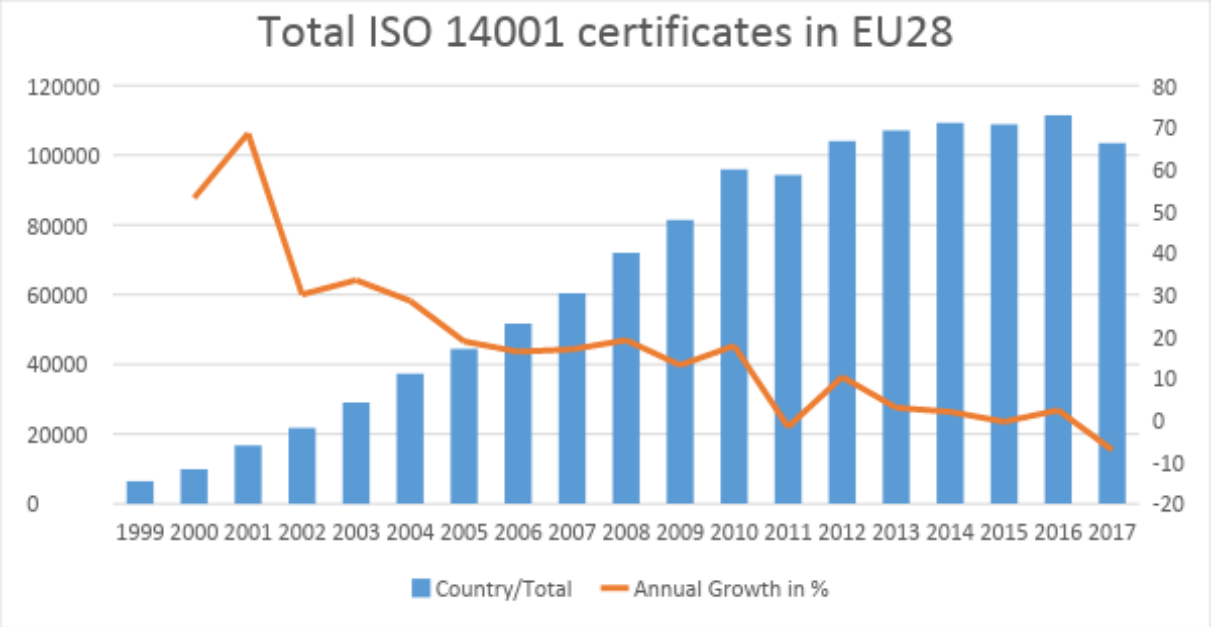


FIG 2: Number of ISO 14001 certificates in EU28

To contrast the diffusion phases of the well established standards on quality and environmental management, we selected two new management standards, which have achieved a certain level of diffusion to identify robust diffusion profiles. According to the ranking based on certificates in Table 1, ISO 27001 on information security management published in 2005 is positioned as third important management standard. However, we did neither select ISO 22000 on food safety management nor ISO 13485 related to medical devices for the pilot, because of the following reason. ISO 22000 is in comparison with ISO 9001 just a more procedural orientated guidance for quality management in the food sector, which can be closely incorporated with the quality management system of ISO 9001. Therefore, this standard is related to the content to close to ISO 9001. ISO 13485 is also harmonized with ISO 9001. In addition, it is often seen as the first step in achieving compliance with European

regulatory requirements, although European Union Directives do not mandate certification to ISO 13465. Therefore, we selected ISO 50001 on energy management, which has only be released in 2011, but shows a rapid and continuous growth.

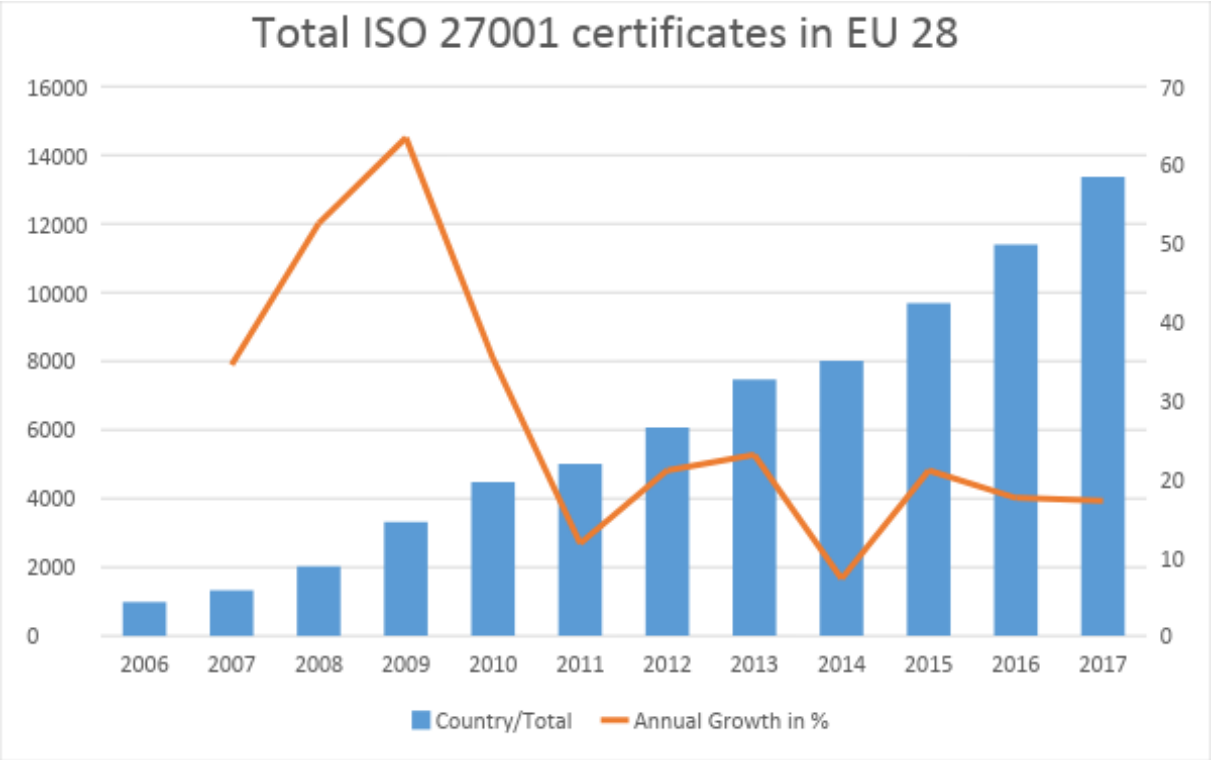


FIG 3: Number of ISO 27001 certificates in EU28

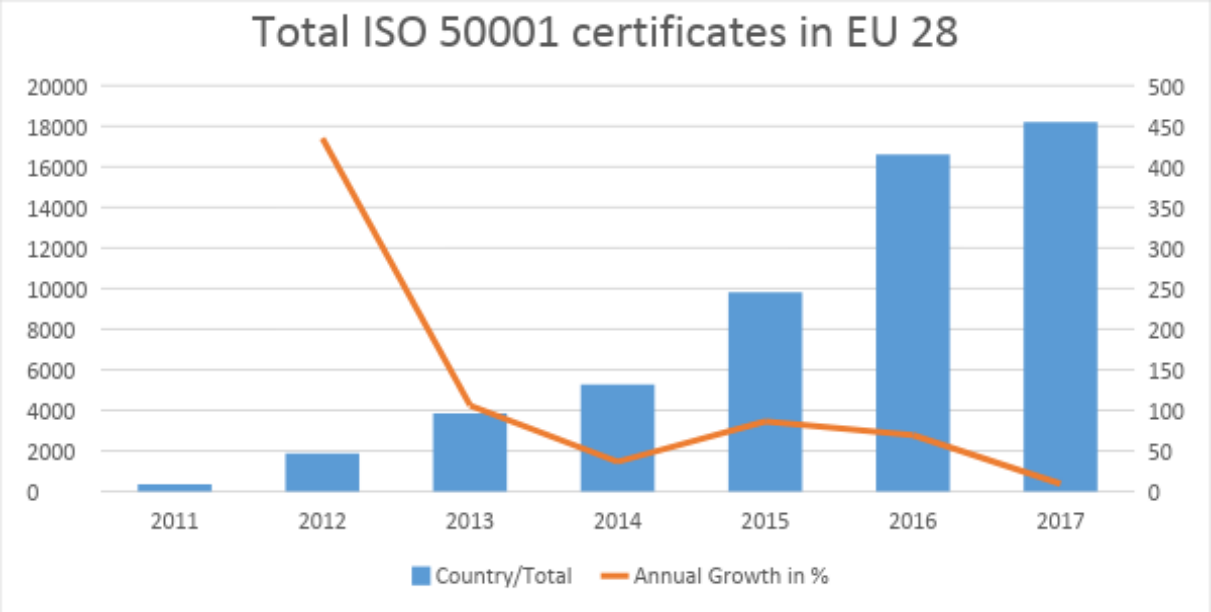


FIG 4: Number of ISO 50001 certificates in EU28

3.2 Sector distribution in EU28

Although, the certification bodies are not able to attribute all of their certified companies to a specific sector, around three quarters can be categorized by sector. The differentiation by sector reveals the relative importance of the various standards for specific industries. The machinery sector is responsible for more than one third of the ISO 9001 certificates. Furthermore, quality management is obviously also relevant for the different types of services confirming e.g. Blind (2005).

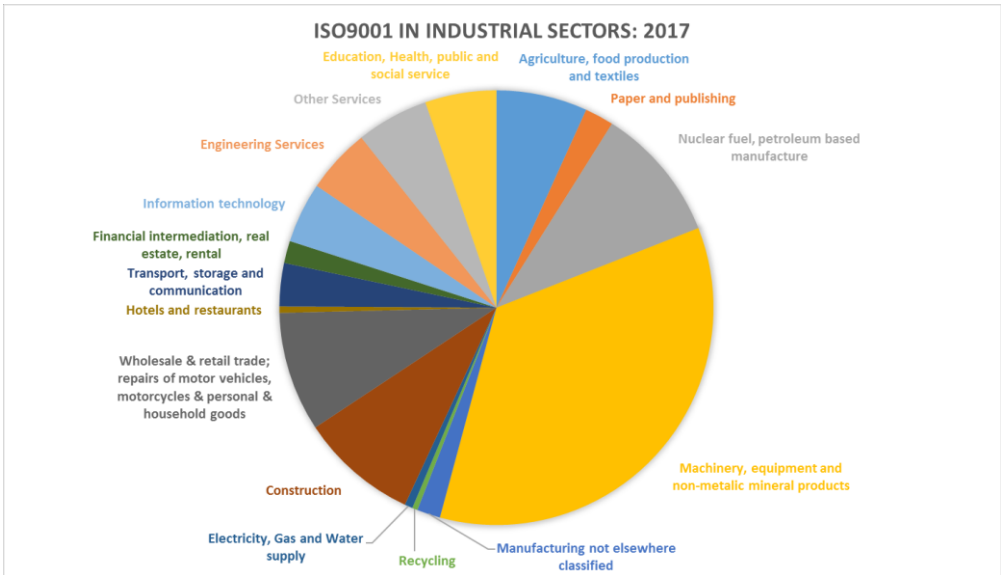


FIG 5: Shares of ISO 9001 certificates in EU28 by sector

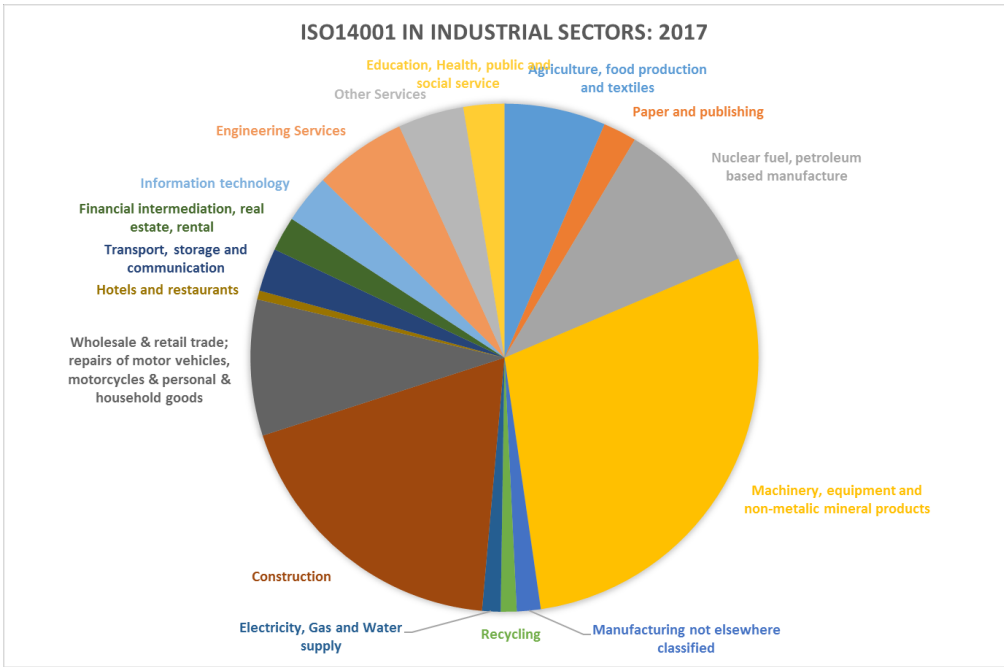


FIG 6: Shares of ISO 14001 certificates in EU28 by sector

The major share of certificates of the environmental management standard ISO 14001 is again taken by companies in the machinery sector, which has also the largest number of companies. However, companies of the construction sector are second in implementing this environmental management standard based on the number of certificates. Furthermore, it is obviously important for the energy sector.

The distribution by sector changes completely in case of ISO 27001, the information security standard. More than half of the certificates is awarded to companies active in information technology, whereas for the machinery sector the topic is not yet important despite the continuing digitalization in the machinery sector.

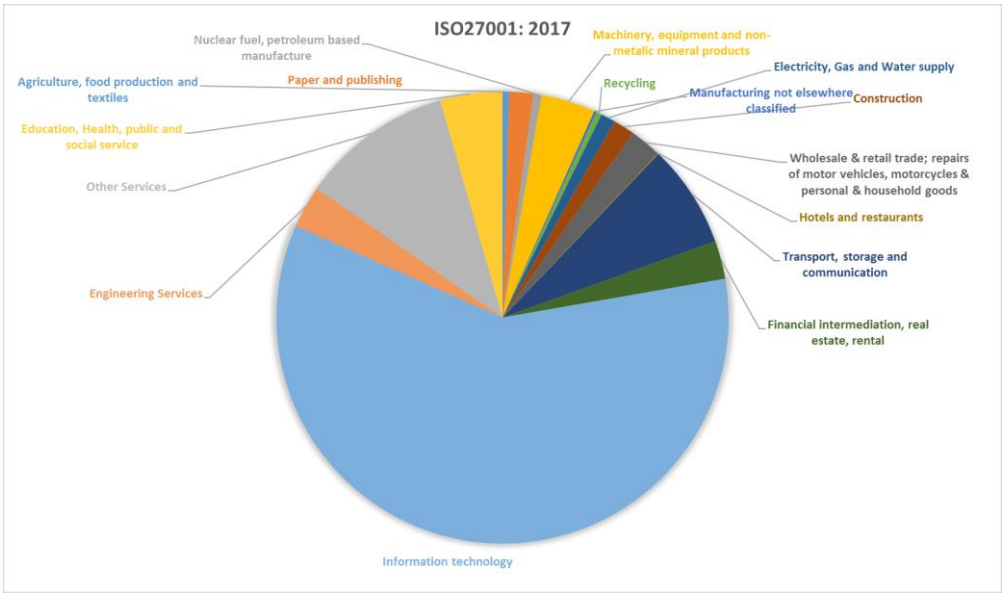


FIG 7: Shares of ISO 27001 certificates in EU28 by sector

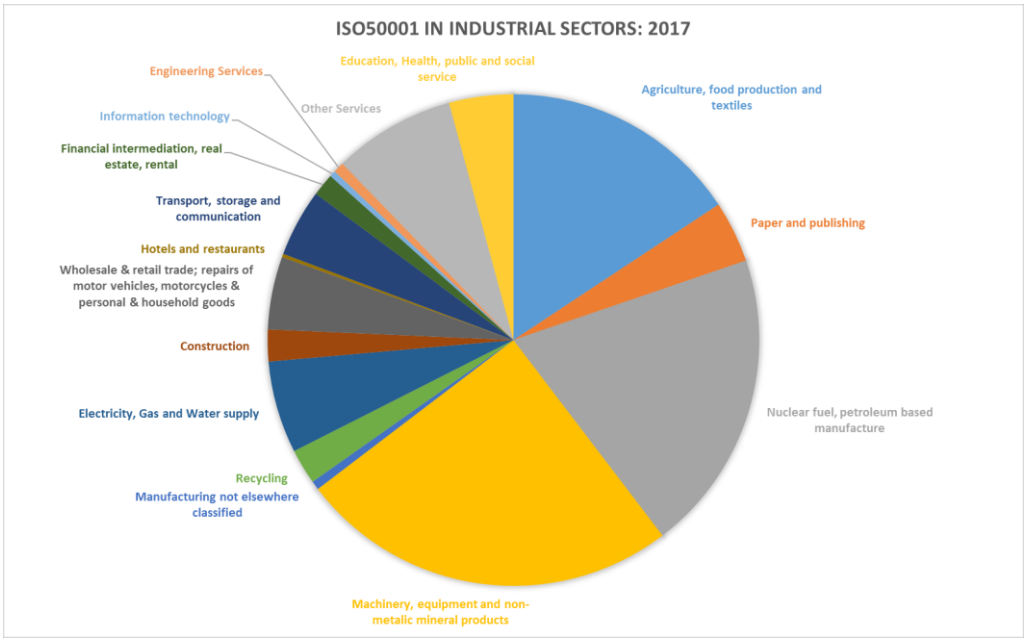


FIG 8: Shares of ISO 50001 certificates in EU28 by sector

Energy efficiency management is not only relevant for the machinery sector, often using energy-intensive production technologies, but also for the energy producing sector and the energy supplying industry.

In summary, the distribution of the four certificates on management standards by industry reflect on the one hand the number of companies in the various industries, but also the industry-specific needs for specific standards, which confirms the validity of the data.

3.3 Certification intensity in EU28 Member States

Most important for the connectivity of certificates as indicator for the diffusion of process innovation to the European Innovation Scoreboard is the development of an indicator, which is applicable for the 28 EU Member States. In contrast to patents and trademarks, which can be applied or registered for many technologies and products of a company, a company is certified or not certified based to a specific standard. Therefore, we use the total number of companies as denominator to construct the indicator certification intensity. Although, we have the certification data for 2017, the number of companies per Member State is in general only available for 2016, for Greece, Ireland, Latvia, Malta and Spain only for 2015 and for Cyprus and Denmark even only for 2014.

Unexpectedly, Italy has the highest intensity of ISO 9001 certificates of around 4% followed by Germany. Furthermore, Romania and Bulgaria have a certification intensity above 1%. All other EU Member States are slightly below 1%.

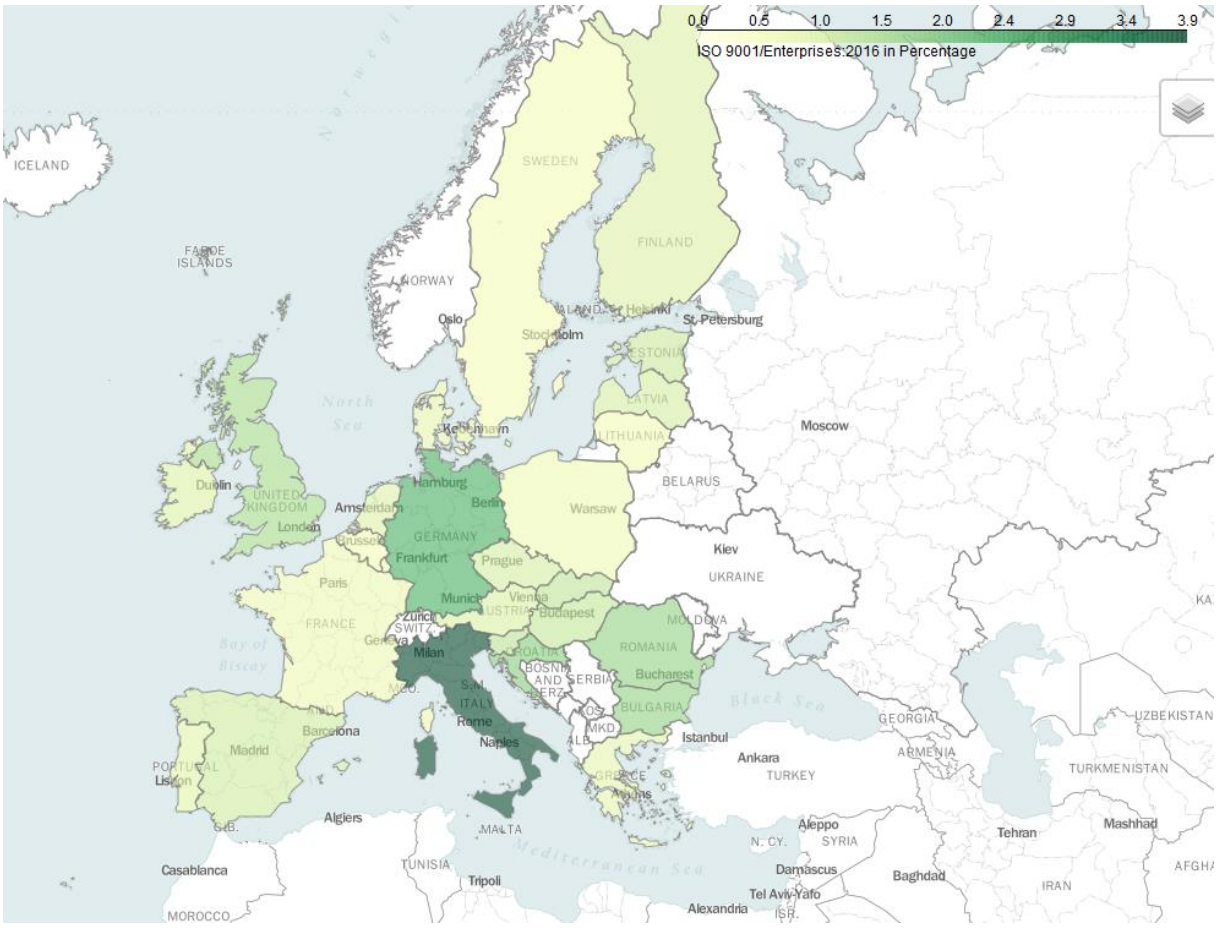


FIG 9: Certification intensity related to ISO 9001 in EU28 by country

The certification intensity is for the ISO 14001 standard consequently much lower in the Member States according to Figure 2. Surprisingly, Romania is leading the ranking in ISO 14001 again followed by Italy and then United Kingdom.

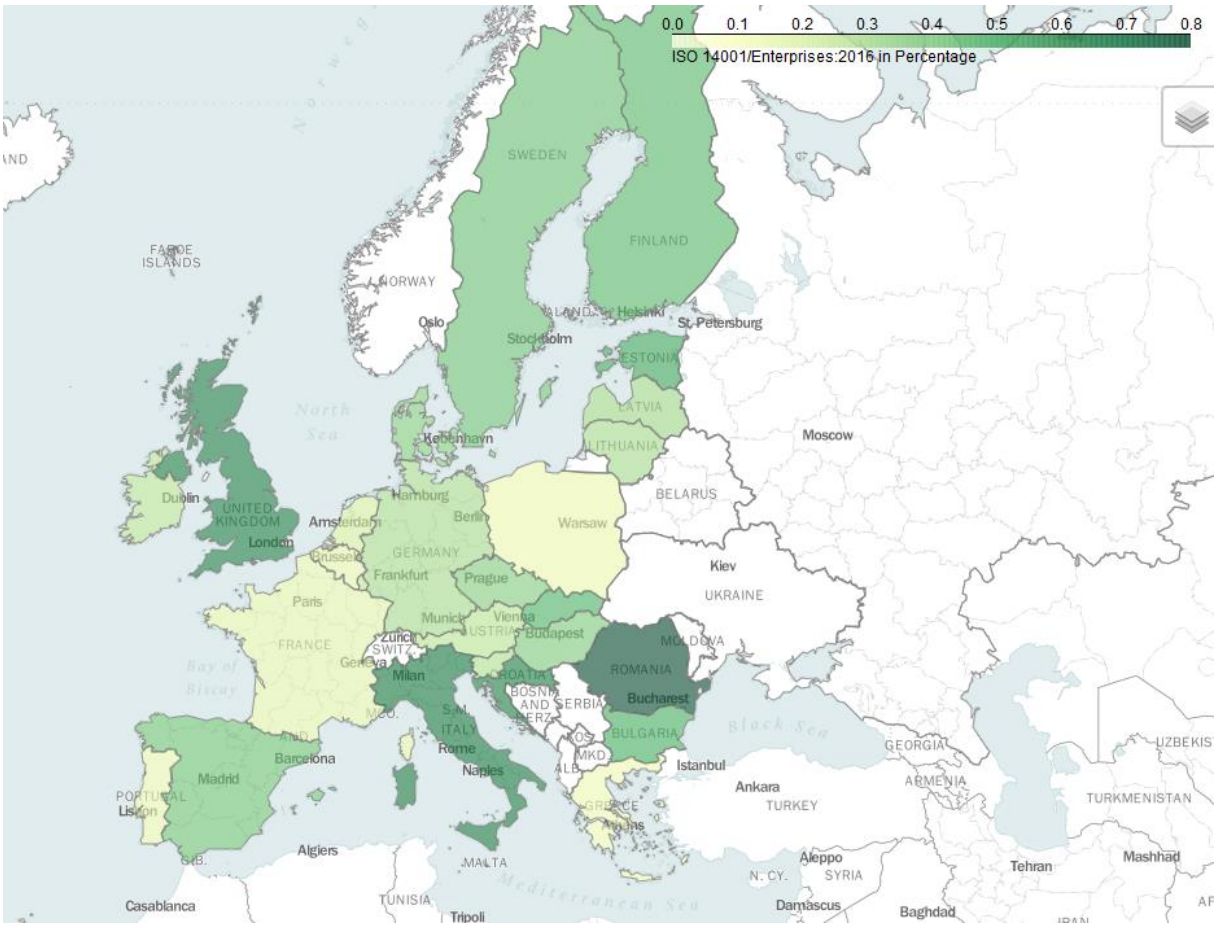


FIG 10: Certification intensity related to ISO 14001 in EU28 by country

However, the United Kingdom has the highest density in ISO 27001 certificates reflecting its relative strength of its information technology sector, which companies' are obviously using this management standard to assure a high level of information security. In addition, it has to be mentioned that this standard has been initially a British standard. Again, we find higher intensities in some of the Member States in Eastern Europe.

Finally, Germany is number one in the ranking of the intensities of the energy management certificates ISO 50001 probably reflecting the change in the energy system, whereas companies in the other Member States in the EU are less interested in getting their energy management certified according to ISO 50001.

Analysing the certification intensity by the Member States of the EU reveals some differences, which can be explained by various factors. First, the sectorial structure of a country has an influence on its certification intensity. For example, the strong manufacturing sector in Italy is contributing to its high intensity in ISO 9001, whereas the leading position of the United Kingdom in the intensity in ISO 27001 can be explained by its strong IT sector. Second, companies in countries with a not yet developed legal system are often using certificates on international management standards for avoiding uncertainty (Orcos et al. 2018). Third, country-specific policies, like the change within the German energy system, towards renewable energy sources have obviously also implications on companies' energy efficiency management, which can be expressed by process innovations, like the introduction of a certified energy efficiency management. Consequently, the number of certificates on international management standards can be considered as an indicator for process innovation for the Member States of the EU.

Based on the intensities in the 28 Member States we calculated correlation coefficients. Whereas the intensities between ISO 9001 and ISO 14001 are highly correlated, the links to the rather new certificates ISO 27001 and ISO 50001 are rather weak. This underlines that the former two are more part of a common institutional framework, to which the latter younger management standards do not belong to yet. They are probably still in a dynamic diffusion process.

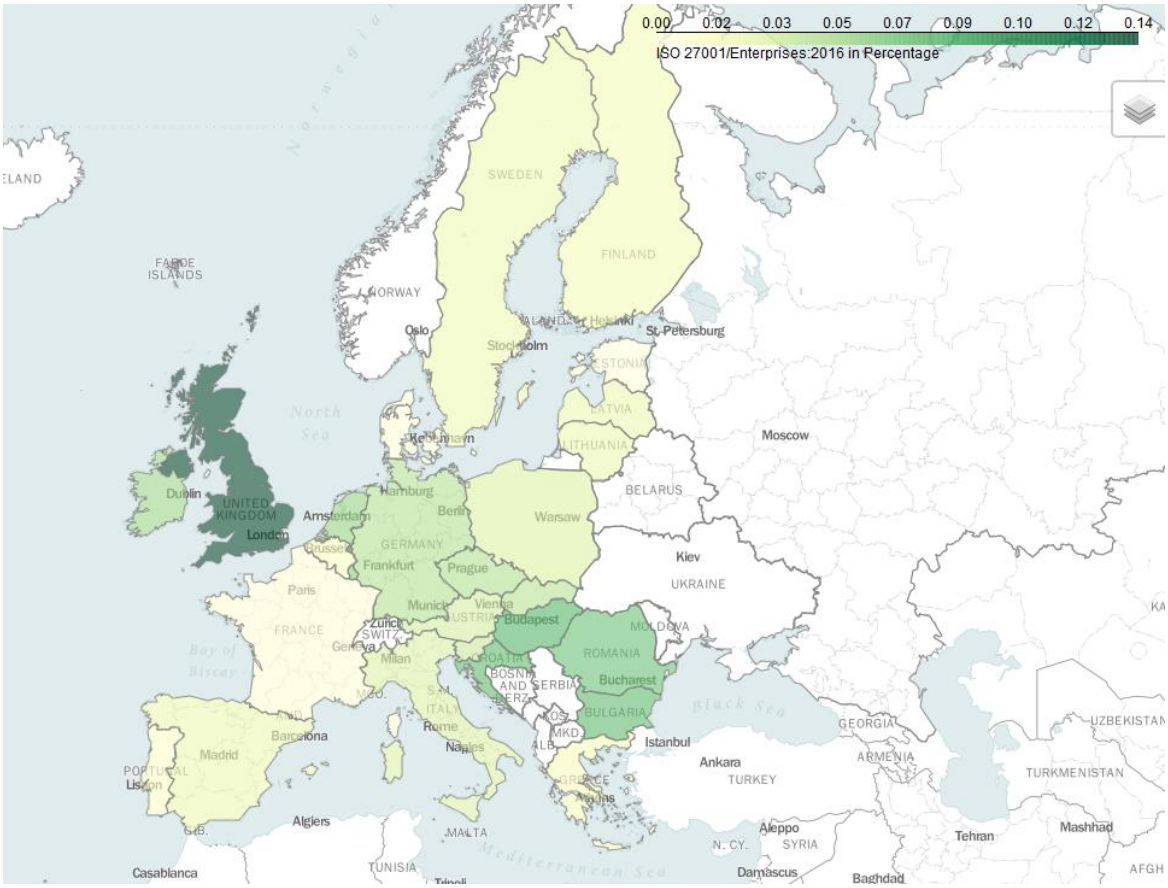


FIG 11: Certification intensity related to ISO 27001 in EU28 by country

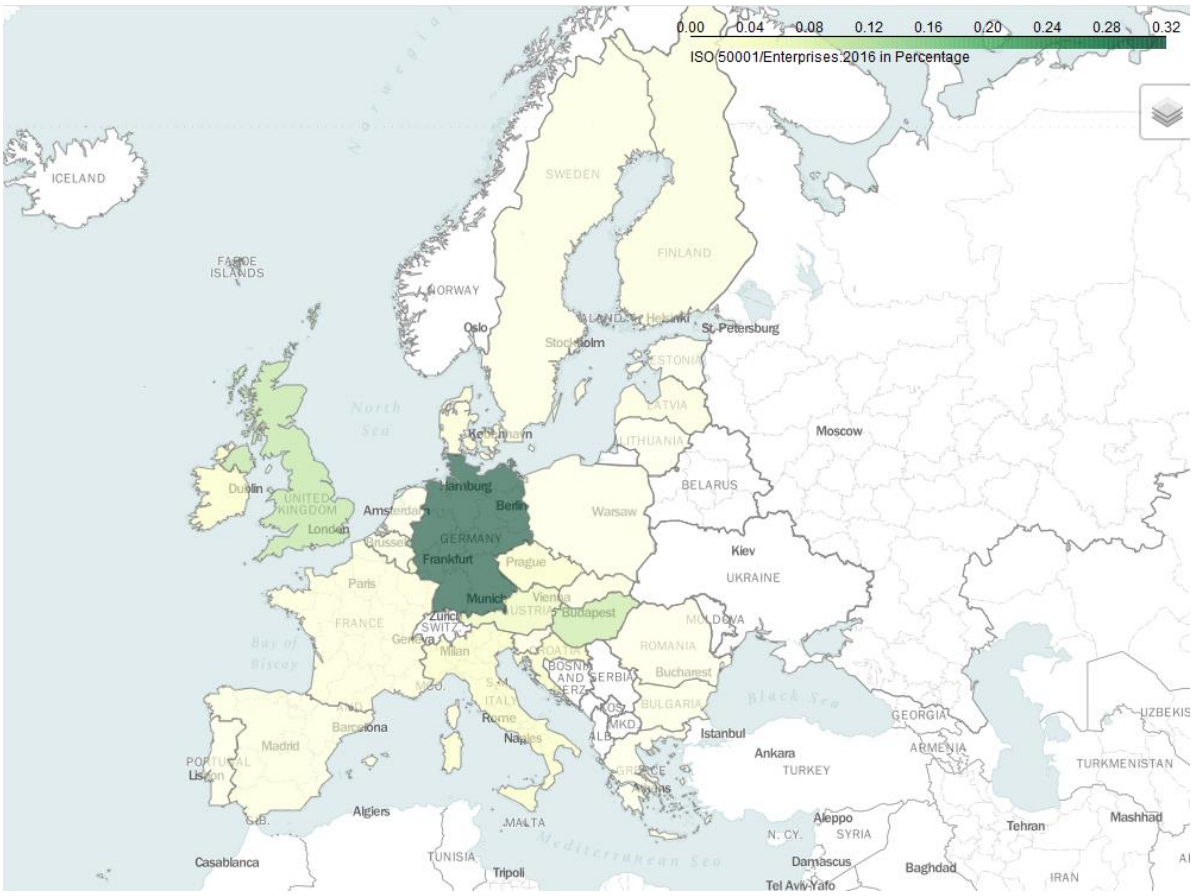


FIG 12: Certification intensity related to ISO 50001 in EU28 by country

3.4 Certification intensity in the context of European Innovation Scoreboard

Certifications related to international management standards can be considered as different elements within the structure of indicators of the European Innovation Scoreboard (European Commission 2017). First, standards are part of the **Framework Conditions** and could be attributed under the subcategory of the *Innovation-friendly environment*, i.e. the number of certifications of management standards as indicator for innovation driving "routines". For example, environmental standards drive environmental innovation as shown by Prakash and Lim (2014). Second, under **Investments** in general and *Firm Investments* in particular, the number of certifications can be an indicator for the costs for getting the certifications as non-R&D expenditures. Third, under the subcategory *Innovators* within **Innovation Activities**, the number of certifications can be used as indicator for process innovations, but also as impulse for further process innovations. Within the other subcategory *Intellectual assets* belonging to **Innovation Activities** certifications could complement other proprietary signalling indicators, like trademarks or design applications.

Guided by these theoretical considerations, we have run correlation analyses based on the data for the 28 Member States of the EU not only for the mentioned indicators used in the European Innovation Scoreboard, but for all of them. In general, the correlation analysis with the indicators of the European Innovation Scoreboard reveals rather low correlation coefficients. Among the framework conditions, only the intensities of ISO 27001 correlate positively with opportunity-driven entrepreneurship. The intensities of the rather new ISO 50001 correlate not only positively with R&D and non-R&D expenditures in the business sector, with the share of Enterprises providing ICT training categorized under firm investments, but also with PCT patent applications and design applications. Finally, it correlates with all impact indicators, i.e. the employment in knowledge-intensive activities and in fast-growing firms in innovative sectors, and all three sales indicators. The intensities of ISO 27001 are also

positively correlated with employment in fast-growing firms and the sales of medium and high tech product exports and of new-to-market and new-to-firm innovations.

Summarizing the insights of the correlation analysis reveals some new insights. In contrast to the theoretically expected correlations, the intensities of certifications are not positively correlated with the indicators for the framework conditions, but in general more with the impact indicators. In addition, the quality and environmental management standards are obviously so well established, that they are not correlated with the innovation indicators used in the European Innovation Scoreboard. Despite this missing correlation at the country level, Rammer et al. (2016) reveal a close link not only between company size, but also companies' innovativeness and their likelihood to certify their quality or environmental management systems. Therefore, certifications of international management standards as indicators for intellectual assets that can be implemented by every company as quality-improving process innovation, but also as environmental innovation focusing e.g. on sustainability.

In summary, certifications of international management standards normalised by countries' number of companies have the potential to be established as an indicator for process innovation. This opportunity responds on the one hand to the recent integration of standards in the 4th edition of the OECD Oslo Manual (OECD 2018). On the other hand, the indicator can be easily integrated as a further complementary indicator for process innovations in the set of indicators used for the development of the European Innovation Scoreboard. In addition, the indicator can also be constructed for the most important countries outside the European Union.

4. Discussion and Conclusions

4.1 Validation

In addition to the internal validation checks and establishing links to the indicators used in the European Innovation Scoreboard, we continued the validation of standards as innovation indicators via the implementation of a webmining approach following the insights by Gök et al. (2015), who reveal that the disclosed innovation related information is more downstream or customer-oriented,. Based on Kinne and Axenbeck (2018) 1.15 million companies out of the 2.52 economically active companies in Germany have an URL. We defined search strings for identifying the four international management standards on companies' homepages applying the webscraper ARGUS (Kinne 2018). According to the ISO Survey, 64.658 companies in Germany are certified based on ISO 9001, i.e. 2.5% of all companies. The webscraping revealed 33.773 companies with websites making a reference to ISO 9001, i.e. 2.9% of the German companies with webpages. For the environmental management standards, the intensity is 0.4% based on the ISO data and 0.48% relying on the webscraping. Furthermore, 8314 certificates related to the energy efficiency management standard ISO 50001 have been awarded to German companies in 2017, i.e. the certification intensity is 0.33%, whereas via webscraping only 2354 companies could be identified with a reference to this standards, which represents an intensity of 0.2%. Here, the webscraping approach leads to a 50% lower intensity. Finally, the information security management standard ISO 27001 as been awarded to only 1339 times in 2017 to German companies according to the ISO Survey, i.e. the certification intensity is only 0.05%. However, the webscraping reveals 2354 companies and therefore a four times higher certification intensity of 0.2%. In summary, the webscraping allowed to identify companies making references to the four selected international management standards, which eventually generated intensities, which are quite similar to the intensities based on the data reported by the ISO survey. The only exception are the data for ISO 27001, which is obviously much more referenced on companies' homepages compared to the issued certificates. However, the webscraping the well established quality and environmental management standard resulted in certification intensities being in the same range than those based on the data provided by

ISO. Therefore, the webscraping approach is not only endorsing the data provided by ISO, but itself is also supported by the ISO data. Obviously, it is possible to identify companies being certified via webscraping, i.e. a new approach to reveal and analyse companies performing process innovations.

The significant differences between the results based on ISO data and webscraping in the case of ISO 27001 motivated us to analyse in detail the webpages with the references to ISO 27001 (Mirtsch and Blind 2019). Preliminary results based on the analysis of a subsample of 300 webpages reveal that indeed almost one half of the companies are certified according to ISO 27001. Another fifth of the companies refer to partner companies, often cloud providers, which are certified according to ISO 27001. This can be characterized as a kind of indirect certification. Finally, one other third of the companies offers consultancy about ISO 27001. Despite the preliminary character of these results, it is obvious that via webscraping it is possible to identify directly or indirectly certified companies. And even the offerings about consulting on ISO 27001 are an indicator about the general relevance or diffusion of this standard in the German economy, because this supply is only developing due to the existence of a significant demand. Overall, these different companies covering both the demand and the supply side related to ISO 27001 is representing its diffusion. Therefore, the aggregate number of companies referencing a specific standard identified via webscraping is an indicator for its diffusion within an economy.

In addition, further information about the identified companies, like sector, size, age and regional location (Kinne and Axenbeck 2018), can be used to identify diffusion patterns. Preliminary analysis show, for example, that the certified companies are significantly younger than the average age of the universe of almost 900.000 companies, for which the company age is available. Although the companies are significantly younger, they are three times as big as the average universe of around a half million companies, for which employment data is available. Finally, but not very surprising, more than one third of the companies belong to ICT services, which confirms the data provided by ISO (see Figure 7). Even, the regional pattern of companies referencing ISO 27001 can be constructed and analysed. Recently, Kinne and Lenz (2019) have developed an innovation indicator based on companies' webpages, which can – in the future – be linked to companies' certification pattern. Overall, this pilot about the certifications related to ISO 27001 within the pilot on standards as innovation indicator reveals some option for further applications of webscraping for the analysis of standards as innovation indicators.

In summary, the pilot analysis confirms that ISO certifications related to international management standards can be used as indicator for process innovations. In addition to quality management, they are addressing several dimensions of sustainability. Since they do not correlate positively with the majority of the current indicators of the European Innovation Scoreboard, they represent obviously an additional innovation dimension, which are closer to process than product innovations. The data is both timely and publicly available. Our initial analysis shows that there is a complete coverage of all EU countries, but also inclusion of the Western Balkan countries and global competitors plus more than additional 100 countries worldwide is possible.

4.2 General limitations

- The data quality of certifications is continuously improved, in particular in some countries, which restricts the validity of some country-specific time series.
- Updates of certification schemes due to new editions of international management standards create frictions in time series.
- The number of companies provided by Eurostat is not updated in all EU Member States to 2017, which challenges the harmonisation of the indicators.

4.3 Ongoing stakeholder involvement

Since standardisation is acknowledged as an innovation activity and standards as an output for research or market integration in the interim evaluation of Horizon 2020 and standardisation has been integrated in the 4th edition of the Oslo Manual (OECD/EUROSTAT 2018), it can be expected that the pilot will contribute to the capacity of standardisation to be moved to the core of R&I policy. In addition to DG Research and Innovation, the currently running Joint Initiative on Standardisation endorsed by DG GROW is interested in the impacts of European standards in general, but also on innovation in particular. Here, Fraunhofer is directly involved in a - not yet published - feasibility study about impacts. In addition, standards are an important element for the integration of the Single Market in general and the Single Digital Market in particular. Further support can be expected by the European standardisation bodies CEN and CENELEC, which have a strong focus on innovation and have recently published an innovation plan aiming to open their standardisation processes and standards to innovators and researchers. Knut Blind is Chairman of a Working Group Standardisation Innovation and Research STAIR, which is collecting data about standards as output of Horizon 2020. ISO created in 2013 the ISO/TC 279 Innovation Management, which was closely linked to the OECD activities related to the 4th edition of the Oslo Manual. Consequently, further support from ISO/TC 279, to which links have been established by Knut Blind during the drafting of the new edition of the Oslo Manual, can be expected. At the national level, the German Ministry of Economic Affairs and Energy have integrated standardisation in their transfer programmes and it is expected that it will play an important role in the currently developed new transfer strategy. Knut Blind is in close exchange with the responsible scientific officer at ministry in preparing the new transfer strategy in order to assure that standardisation is adequately integrated.

4.4 Considerations for scaling up

4.4.1 Complementarities with other pilots

Standardisation activities in Artificial Intelligence and the Bioeconomy have been considered to be considered in Pilot 1 on Emerging Tech Ecosystems being run by Nesta and Pilot 3 on structural technological change being performed by DTU. Standardisation could be also relevant for the mission-oriented pilot.

4.4.2 Scaling up

Since the pilot is based on international standards, it is already positioned at the European or even international level. The pilot could be easily scaled up to include all ten international However, the option of generating further insights from extracting information about companies' certifications related to international management standards from firm websites of other EU Member States in addition to Germany using web scraping is in general more promising. It could contribute to the validation of the web scraping approach performed in Germany.

Unfortunately, the webscraping of all companies within the other EU Member States is not feasible, because alone the identification of companies' webpages requires resources, which are not available. Furthermore, information about companies' references to international management standards or even other standards is not sufficient. Company specific data, e.g. about size, sector and age, is needed to put standards as an innovation indicator into context. Therefore, it is proposed to focus the web mining on the 577 European companies within the sample of the EU Industrial R&D Investment Scoreboard (http://iri.jrc.ec.europa.eu/scoreboard18.html), because here company information is available to explain their referencing and use of standards in general, not only the selected international management standards. In consequence, the feasibility of extending the applied approach to standards in general will be tested.

4.4.3 Tools

The feasibility of scaling up the web scraping in other EU Member States has to be checked and might require additional external resources, which currently cannot be specified.

5. References

Acemoglu, D.; Gancia, G.; Zilibotti, F. (2012): Competing engines of growth: Innovation and standardization. Journal of Economic Theory 147 (2012), 570–-601.

Akerlof, G. A. (1970): The Market for "Lemons": Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics , 84(3), 488-500.

Arthur, W. B. (1989): Competing Technologies, Increasing Returns, and Lock-In by Historical Events. The Economic Journal, Vol. 99, No. 394, 116-131.

Blind, K. (2006): A Taxonomy of Standards in the Service Sector. The Service Industries Journal 26(4), 397-420.

Blind, K. (2019a): Standardization and Standards as Science and Innovation Indicators, forthcoming in: Springer Handbook of Science and Technology Indicators.

Blind, K. (2019b): Standardization in the Context of Evolutionary Economics, forthcoming in: Handbook of Research Methods and Applications in Industrial Dynamics and Evolutionary Economics

Blind, K.; Mangelsdorf, A,; Niebel. C.; Ramel, F. (2018): Standards in the global value chains of the European Single Market. Review of International Political Economy 25 (1), 28-48.

Blind, K.; Mangelsdorf, A.; Pohlisch, J. (2018): The effects of cooperation in accreditation on international trade: Empirical evidence on ISO 9000 certifications, in: International Journal of Production Economics. 2018, 198, 50-59

Blind, K.; Petersen, S.; Riillo, C. (2017): The Impact of Standards and Regulation on Innovation in Uncertain Markets. Research Policy 46 (1), 249–264.

Clougherty, J.A. & M. Grajek (2008): The impact of ISO 9000 diffusion on trade and FDI. A new institutional analysis. J. Int. Bus. Stud. 39 (4), 613–633.

Clougherty, J.A. & M. Grajek (2014): International standards and international trade. Empirical evidence from ISO 9000 diffusion. Int. J. Ind. Organ. 36, 70–82.

European Commission (2017): Intervention logic of Horizon 2020 interim evaluation, https://ec.europa.eu/research/evaluations/pdf/archive/h2020_evaluations/intervention_logic_h2020_052016.pdf#view=fit&pagemode=none

European Commission (2018): European Innovation Scoreboard 2018, https://ec.europa.eu/docsroom/documents/33147

Gawer, A. (2014): Bridging differing perspectives on technological platforms: Toward an integrative framework. Research Policy 43 (7), 1239-1249

Geels, F. (2002): Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. Research Policy 31(8–9), 1257-1274.

Geroski, P. (2000): Models of technology diffusion. Research Policy, 2000, vol. 29, issue (4-5,), 603-625.
Gök, A.; Waterworth, A.; Shapira P. (2015): Use of web mining in studying innovation. Scientometrics 102 (1), 653-671.
Guler, I.; Guillen, M.; Macpherson, J. (2002): Global Competition, Institutions, and the Diffusion of Organizational Practices: The International Spread of ISO 9000 Quality Certificates. Administrative Science Quarterly, 47 (2), S. 207-232.
International Organization for Standardization (ISO) (2018): The ISO Survey. Retrieved from http://www.iso.org/iso/iso-survey
Kinne, J (2018): ARGUS: Automated Robot for Generic Universal Scraping. Repository: https://github.com/datawizard1337/ARGUS
Kinne, J; Axenbeck, J. (2018): Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany, ZEW Discussion Paper No. 18-033, Mannheim.
Kinne, J; Lenz, D. (2019): Predicting Innovative Firms using Web Mining and Deep Learning. ZEW Discussion Paper No. 19-001, Mannheim.
Levine D. I & M. W. Toffel (2010): Quality Management and Job Quality: How the ISO 9001 Standard for Quality Management Systems Affects Employees and Employers. Management Science 56(6):978-996.
Lim, S. & A. Prakash (2014): Voluntary Regulations and Innovation: The Case of ISO 14001. Public Administration Review, 74, 233–244.
Mirtsch, M. & K. Blind (2019): Diffusion of the Information Security Management System Standard ISO/IEC 27001: An explorative Web Mining analysis, Paper submitted to EURAS 2019.
Nelson R.R.; Winter S.G (1982): An evolutionary theory of economic change. Harvard Univ. Press. Cambridge, MA.
OECD/Eurostat (2018): Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg.
Orcos, R.; Pérez-Aradros, B; Blind, K. (2018): Why does the diffusion of environmental management standards differ across countries? The role of formal and informal institutions in the adoption of ISO 14001. Journal of World Business 53 (6), 850-861.
Pohle, A.; Blind, K.; Neustroev, D. (2018): The Impact of International Management Standards on Academic Research. Sustainability 2018, 10, 4656
Potoski, M. & A Prakash (2014): Do voluntary programs reduce pollution? Examining ISO 14001's effectiveness across countries. Policy Studies Journal 41 (2), 273-294
Rammer, C.; Berger, M.; Doherr, T.; Hud, M.; Hünermund, P.; Iferd, Y.; Peters, B.; Schubert, T. (2016): Innovationsverhalten der deutschen Wirtschaft - Indikatorenbericht zur Innovationserhebung 2015, Bundesministerium für Bildung und Forschung (Hrsg.), Mannheim: ZEW.

Robertson, T.S. & H. Gatignon (1986): Competitive effects on technology diffusion. Journal of Marketing 50, 1–12.
Rogers, E. M. (1995): Diffusion of innovations (4th Edition). New York: Free Press.
Rysman, M.; Simcoe, T.; Wang, Y. (2018): Differentiation in Adoption of Environmental Standards: LEED from 2000-2010, under revision for Management Science.
Schumpeter, J. A. (1939): Business Cycles. A Theoretical, Historical and Statistical Analysis of the Capitalist Process, New York/London.
Spence, M. (1973): Job Market Signaling. The Quarterly Journal of Economics, 87(3), 355-374.
Spring, M. B., Weiss, M. B. (1995): Financing the standards development process. Standards Policy for Information Infrastructure, 289-320.
Suarez, F. F. (2004): Battles for technological dominance: an integrative framework. Research Policy 33, 271-286.
Swann, P. (2000): The Economics of Standardization. Final Report for Standards and Technical Regulations Directorate Department of Trade and Industry
Tassey, G. (2000): Standardization in technology-based markets. Research Policy, 29 (4-5), 587-602.
Terlaak, A.; King, A. A. (2006): The effect of certification with the ISO 9000 Quality Management Standard: A signaling approach. Journal of Economic Behavior & Organization 60 (4), 579-602.
Tuczek, F.; Castka, P.; Wakolbinger, T. (2018): A review of management theories in the context of quality, environmental and social responsibility voluntary standards. Journal of Cleaner Production 176, 399-416.
Utterback, J. M.; Abernathy, W. J. (1975): A dynamic model of process and product innovation. Omega, 3(6), 639-656.
Wakke, P.; Blind, K.; Ramel, F. (2016): The impact of participation within formal standardization on firm performance. Journal of Productivity Analysis, 45(3), 317-330.
Weitzel, T.; Beimborn, D.; König, W. (2006): A Unified Economic Model of Standard Diffusion: The Impact of Standardization Cost, Network Effects, and Network Topology. MIS Quarterly, 30, 489-514.

Pilot 5: Evidence Base For Mission-Oriented Research & Innovation

Abstract

Mission-driven innovation policies to address specific technological, social or economic challenges are being increasingly recognised as a valid strategy to steer innovation in societally desirable directions and encourage the formation of new disciplines and industries. The formulation of this policy agenda has however raced ahead of the evidence base, creating the risk that mission-driven innovation policies are insufficiently informed in their formulation, targeting, monitoring and evaluation. We argue that developing a suitable evidence base for mission-driven policies will require the use of new data sources, analytical methods and indicators that reflect the rationale and pathways to impact of mission-driven policies through network-building and interdisciplinary crossover. We propose an approach that decomposes mission statements into simpler elements that can be used to query innovation databases in order to map potential and active 'mission fields', and develop prototype indicators. We implement this framework in the empirical setting of the UK's grand challenge to "*Use data, Artificial Intelligence and innovation to transform the prevention, early diagnosis and treatment of chronic diseases by 2030*" using data about research funding in the UK. Having presented and discussed emerging findings from this analysis, we consider opportunities to scale up the methodology developed in this pilot in subsequent stages of the EURITO project.

1 Introduction

Innovation policymakers are developing new frameworks for Research and Innovation (R&I) based on the idea of missions: big societal challenges that will be addressed by combining the knowledge from different disciplines in ambitious projects and programmes. This pilot produces maps and metrics to inform the design, implementation and evaluation of mission-oriented research and innovation in the EU (Mazzucato, 2018a, 2018b).

1.1 Background/Context

Opportunities and challenges for a new wave of mission-driven innovation policies

Mission-oriented policies can be defined as systemic public policies that draw on frontier knowledge to attain specific goals or "big science deployed to meet big problems" (Mazzucato, 2018). 'Missions' are at the core of new €100bn EU proposal for Horizon Europe (released June 2018), and of the new industrial strategy developed by UK government, which is organised around the idea of 'Grand Challenges' that generate missions, as well as an 'Industrial Strategy Challenge Fund' that identifies particular challenges to be pursued through concerted policy action (HM Government, 2018). Some mission ideas and policies include "*Reach net zero greenhouse gas emissions balance of 100 European cities by 2030*" (Mazzucato, 2018), or "*Ensure that people can enjoy at least 5 extra healthy, independent years of life by 2035, while narrowing the gap between the experience of the richest and poorest*" (BEIS, 2018). The figure below illustrates a mission specification from Mazzucato (2018).

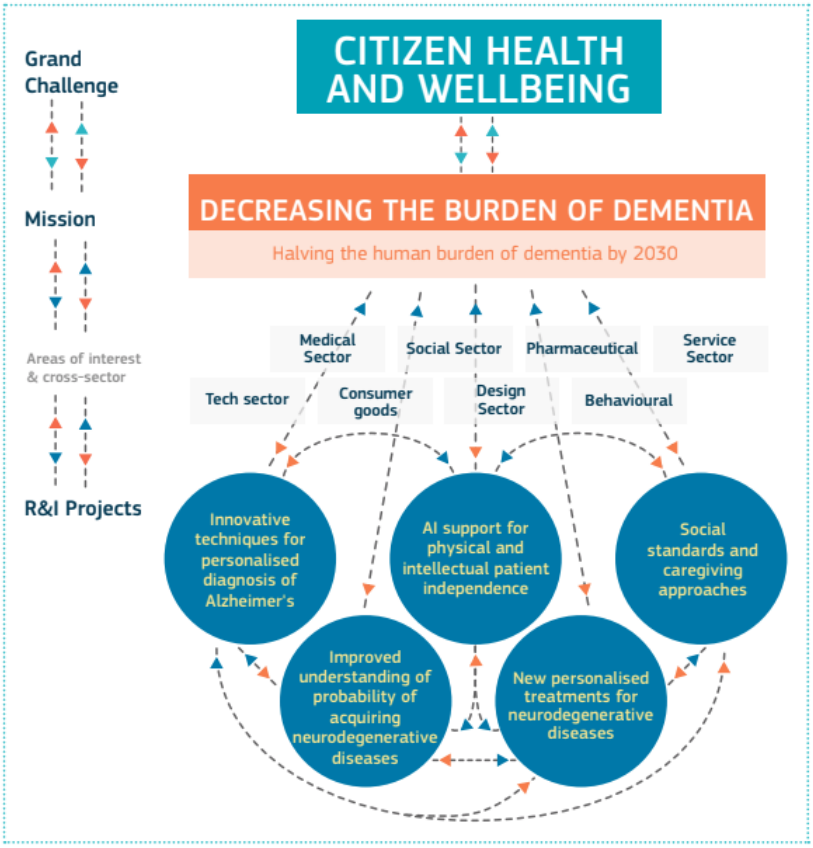


Figure: Example mission (Mazzucato 2018, p. 26)

Missions are not new in R&I policy. Some of humankind's greatest technical achievements have resulted from missions. Some examples include the Longitude Reward, which encouraged the development of sea watches greatly improving the safety of sea travel in the 18th Century, the Manhattan Project to develop an atomic bomb, or the Apollo program to put a man on the moon (Mowery, Nelson, & Martin, 2010; Nelson, 1977). Mission-like policies also play an important part in the private sector through crowdsourcing challenges and competitions (a comparative assessment of some of the biggest and most influential frameworks can be accessed here).

One thing that sets the 'new wave' of missions in R&I policy apart from previous ones is their shift away from purely technical or economic problem-solving, and their ambition to deploy science and technology to address big social challenge that people face in their daily lives. One implicit goal of this effort is to build popular support for R&I policies that can otherwise feel removed from everyday aspirations and concerns about the environment, health, education and inequality (Mazzucato, 2018b). This move away from technical goals to social ones creates new challenges for R&I policy and its stakeholders, as the goal of innovation research and policy shift from advancing knowledge and producing technical breakthroughs to achieving social changes which are harder to measure and in some cases even contested, as we see in ongoing debates about how to tackle climate change or increase social mobility (Nelson, 1977, 2011). This broader and arguably more ambitious definition of missions also requires engaging a wider set of disciplines, overcoming barriers to interdisciplinary collaboration.

Another important reason for renewed interest in the mission framework is the notion that traditional R&I policies based on market failure and system failure rationales ignore the directionality of technical change, have low additionality and tend to be captured by the status quo (Cantner & Vannuccini, 2018; Frenken, 2017). In other words, they support incremental activities that might not be the most societally

desirable, and in some cases would have been carried out anyway, generally by powerful, established incumbents (Frenken, 2017). This way of thinking about the goals of R&I policy draws on evolutionary economics and complexity science ideas arguing that technological development is directional (it can unfold in multiple possible trajectories) and uncertain in outcomes (the trajectory that is selected 'from the bottom up', for example by market forces might not be the most beneficial one because it generates unanticipated externalities and path dependencies that prevent shifts away from it further down the line, as we see in the economy's lock-in to environmentally unsustainable fossil fuels) (Aghion, David, & Foray, 2009; Arthur, 2009; David, 1985). Further. 'normal' R&I takes place through an exploration of the 'adjacent possible' where bodies of knowledge that are technically closer tend to be recombined more often because they exist in the same organisation, or in related organisations (Boschma, 2005; Hidalgo et al., 2018). All this means that R&I policies that simply seek to increase the amounts invested in R&D (as R&D tax credits do under a market failure rationale) or the responsiveness of academic researchers to industry needs (as knowledge exchange programs informed by a systems failure rationale do) will fail to generate societally beneficial, boundary spanning innovations (Gustafsson & Autio, 2011). Recent criticisms of mainstream science policy, research funding in the biomedical domain, and the emphasis of AI research on automating labour instead of complementing it, and evidence of a productivity puzzle where increasing investments on R&D fail to produce corresponding scientific discoveries or improvements in productivity growth are consistent with this view (Acemoglu & Restrepo, 2018; Jones & Wilsdon, 2018; Restrepo & Acemoglu, 2018; Sarewitz, 2016).

A mission-driven R&I framework could help address these challenges: by definition, missions are directional. They identify preferred trajectories of technological development (in the examples we provided at the beginning, use of urban infrastructure and built space innovations to reduce carbon emissions, or of AI to address chronic diseases) and provide resources for pursuing them. They also involve combinations of knowledge residing in faraway parts of the innovation system. This should yield new ideas for which there is not a market yet (otherwise they would already be deployed to address the mission) (Autio, 2011). This could involve new players sitting in the intersection between disciplines, and with a greater tolerance for risk. As before, the ambitions of this new approach come with their own challenges, such as the risk that R&I policymakers without sufficient information end up 'picking winners' that are not feasible or commercially sustainable, or that they are captured by new constituencies that coalesce around missions (Aghion et al., 2009).

Designing missions that work

Proponents of the new wave of mission-driven policies set out to address the design and implementation challenges outlined above by putting in place new criteria for mission selection of delivery (see table 1) (Mazzucato, 2018b).

Table 1: Mission selection criteria, evidence base and indicators

Mission selection criterion	Feature of the evidence base	Indicators
1. Social relevance. Missions should be bold, inspirational, with wide societal relevance	Measure social relevance and engagement with the mission	Alignment of mission objective and expressions of societal need

		Social media engagement with the mission
2. Feasible distinctiveness: Ambitious but realistic R&I activities	Measure the extent to which the activities being supported draw on but are qualitatively different from previous research.	Distinctiveness/relatedness between mission activities and the status quo
3. Induces crossover Cross-disciplinary, cross-sectoral, and cross-actor	Measure if the mission is encouraging new combinations of disciplines, technologies and industries	Disciplinary diversity in missions compared to status quo Novelty of stakeholder groups compared to missions
4. Diverse solutions Multiple, bottom-up solutions.	Measure the diversity of options that are being explored as part of the mission.	Diversity of approaches in the mission field
5. Measured Clear direction: measurable, time-bound	Measure if the R&I activities taking place as part of the mission are achieving its goals	Design features of the mission (KPIs, duration) Attribution of impacts to mission activities

These criteria seek to ensure that missions have legitimacy and broad social support, can be achieved, and bring together a broad mix of actors going beyond 'the usual suspects'. The demand for bottom-up experimentation acknowledges uncertainty about which of the avenues that are explored through the mission will be successful, and to avoid the risk of picking winners that end in a technological or commercial dead-end.

1.1.1 Opportunity

Effective design, delivery, monitoring and evaluation of missions requires a suitable evidence base. We draw on the criteria above to sketch some features of this evidence base, and identify new opportunities to develop new indicators to make it operational (see table 1):

First, an evidence base for missions should capture the extent to which the topic of a mission reflects societal interests and concerns, as well as the levels of social engagement with the mission topic, and with the mission itself. It is possible to generate indicators capturing the social relevance dimension of missions with data from opinion surveys, policy debates, news media and social media.

Second, the evidence base needs to consider the content of the R&I activities taking place as part of the mission, and how they balance feasibility (the activities taking place need to be technologically

plausible and possible) and ambition (the activities would not have taken place without the mission). This can be measured with indicators reflecting the semantic similarity (or distance) between R&I activities taking place as part of the mission and previous work, as well as the extent to which the projects taking place inside the mission generate technological outputs (i.e. are technologically feasible).

Third, it is also important to consider the diversity of activities that are being supported through the mission in terms of the disciplines, industries and actors involved: are these novel and unexpected, and do they involve new 'entrants'? Here we can develop indicators measuring the disciplinary and industry mix of the activities supported by a mission, and compare the actors participating in it with those involved in areas outside of the mission. Over time, we would expect successful missions to change the structure of R&I networks, bringing key disciplines and the actors involved in them into new fields and communities. We can develop indicators that capture this.

Finally, the evidence base should capture the timelines and goals for the mission: what does it seek to achieve, over what period, and with what success. Some of the relevant indicators can be directly extracted from the specification of the mission (eg. reach zero greenhouse emissions by 2030). This is a different question from whether these indicators are being measured in a suitable way, and from the extent to which changes in those indicators (impacts) can be attributed to the mission itself (as compared to broader socio-economic trends and technological breakthroughs supported outside of the mission). This will require experimental designs that will depend on the nature of the mission, its goals and available data.

About

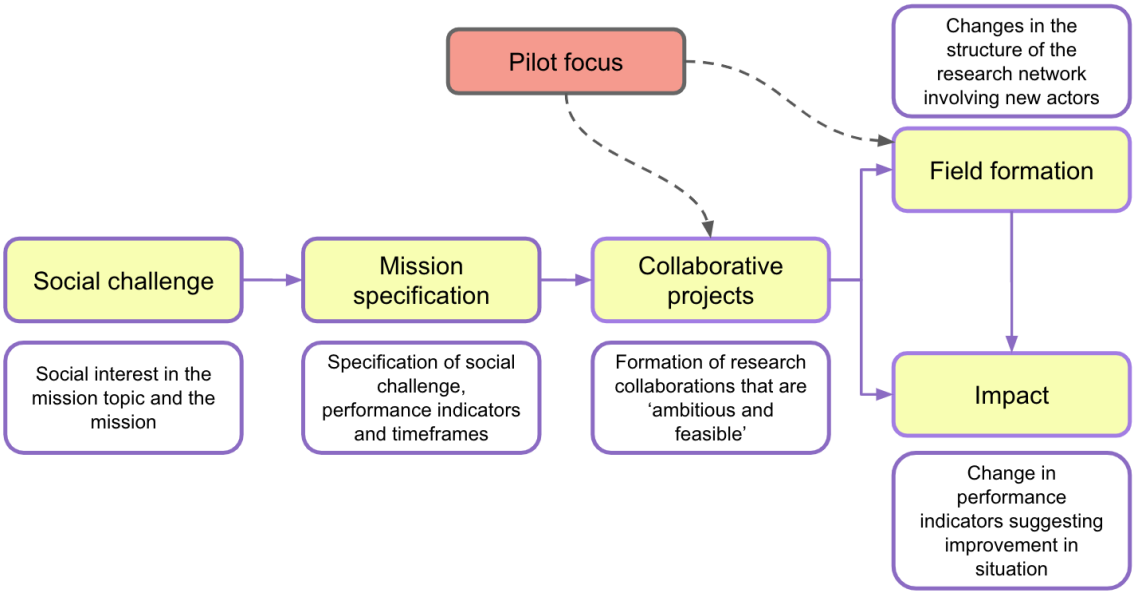
This pilot explores potential indicators to operationalise key dimensions of the evidence base for missions outlined above. We do this in the context of the UK Grand Challenge Mission to "*Use data, Artificial Intelligence and innovation to transform the prevention, early diagnosis and treatment of chronic diseases by 2030*", and using data about research funding in the UK from the Gateway to Research open dataset.

Our goal is to develop a flexible framework that can be used to query funding data and identify a "mission field". We define a mission field as the set of R&I activities that are directly relevant for a mission. We establish relevance using a semantic approach that extracts keywords from the definition of a mission and then queries the funding corpus with a version of that keyword set expanded via semantic similarity in a word embedding space. We then calculate indicators in the mission field. All these activities are described in further detail in the Methodology and Findings sections below.

We focus on dimensions 2, 3, and 4 in Table 1, considering actors, activities and networks but not impacts and social media activity (see figure below for a summary). The reason for this is that mission-driven innovation policies are a novel policy concept and most missions (including the one we are focusing on in this pilot) have only been recently implemented. It would be unrealistic to expect them to have produced visible impacts so early after launch. We consider some options for this (challenging) form of mission-driven policy monitoring and evaluation in the conclusions. At the onset of the pilot we decided to concentrate on the GtR data, which is why we have not used social media data in the analysis.

A next step for us will be to integrate it into our pipeline, possibly through the CrossRef Event API, which allows querying of research paper Document Object Identifiers (DOI) to identify social media activity around them (Ortega, 2018).

The previous point about the nature of the phenomenon we are capturing at this point in time also applies to the actor, activity and outcome indicators that we are focusing on in the pilot - that is to say, the levels of activity we are capturing in our analysis should be seen as a 'baseline' for the mission rather than evidence of its impact, given the short time that the mission has had to change the orientation and configuration of R&I activities.



1.1.2 Application domain

Our application domain are publicly funded R&I activities in the UK related to the application of Artificial Intelligence (AI) and data to the prevention, diagnosis and treatment of chronic diseases. The idea underlying this mission is that general purpose AI (and machine learning) technologies with high predictive potential could be deployed to transform how we deal with chronic conditions such as cardiovascular diseases, cancer or diabetes (Cockburn, Henderson, & Stern, 2018; Loder & Nicholas, 2018).[1] The range of applications is broad, going from identifying the causes of these diseases, to predicting what individuals are at risk, and designing more effective personalised treatments for them (under the rubric of precision medicine). Ultimately, this will contribute to saving lives, improving wellbeing, and lowering costs in healthcare delivery by reducing the need for costly late-treatments (BEIS, 2019).

Why is a mission needed in this domain? There is a general perception that applications of AI in the health domain are lagging behind other application areas such as advertising, social media or finance (J. Mateos-Garcia, 2017; J. C. Mateos-Garcia, 2018; Mulgan, 2017). There are multiple reasons for this including risks of prediction failure, patient data protection and privacy issues, the importance of model explainability and barriers to deployment in large and complex health systems. Overcoming these barriers to the successful deployment of AI in health requires new combinations of knowledge and innovation actors that this mission seeks to encourage.

[1] The World Health Organisation defines chronic diseases as those that "*are not passed from person to person. They are of long duration and generally slow progression. The four main types ... are cardiovascular diseases (like heart attacks and stroke), cancers, chronic respiratory diseases (such as chronic obstructed pulmonary disease and asthma) and diabetes.*"

1.1.3 Flexibility of application domain

How scalable is our approach to other data sources and missions?

In terms of data sources, our approach requires two main inputs: a list of keywords describing the contents of a mission (its domain, purpose, technology etc.) and a list of text descriptions about potentially relevant R&I activities that we query with these keywords. These two inputs are sufficient to map a mission field. Additional metadata such as funding year, level of funding, name of the funder, organisations collaborating in funded projects and outcomes and impacts can generate additional indicators. All this information, is generally available from open datasets about research funding such as the Gateway to Research (which we are using here), the CORDIS database of EU H2020 funded projects,or the National Institutes of Health World Reporter database of health-related research. It is also possible to use this approach in other potentially relevant datasets such as patent databases, pre-prints or open source software (with the caveat that they will generally contain a narrow set of disciplines than what is available in broad-based research funding databases). All this means that the approach that we follow in this pilot is relatively easy to scale up to other data sources.

In principle, the framework that we develop should also be applicable to other missions as long as activities relevant to them are captured in research funding databases, and it is possible to extract keywords from their definition. The facility for this depends on the mission in question. For example, the mission that we selected for this pilot has a relatively well defined set of keywords around 'AI' and 'chronic conditions' that can be mapped on the concepts of 'solution space' and 'problem space'. This makes it easier to delineate and analyse the mission field.[2] Other missions, such as the previously mentioned EU mission to reach zero emissions in EU cities by 2030 would require additional background research to identify bodies of research and knowledge that might contribute to achieving the goal in the context of urban infrastructures (the solution space). Of course, the results of the analysis will be sensitive to the keywords that are selected.

Our decision to exclude mission impacts from our indicator framework also makes our approach more scalable, since that is one dimension where missions are likely to be highly heterogeneous. One could think of the framework that we present here as a modular component of a broader measurement framework also including indicators of R&I impact that would be specific to the mission under analysis.

1.1.4 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

There are several ways in which this pilot will contribute to incorporate currently peripheral measurement considerations into the mainstream of innovation policy:

Our indicators capture directional considerations: they relate to specific technological applications and socio-economic challenges that have been identified as target priorities in R&I policy. This is very different from the current portfolio of indicators, that approach innovation investments as homogeneous and fungible independently of their purpose.

Our indicators pay particular attention to dimensions of diversity within a mission field, including the variety of disciplines and industries that are involved, and the variety of technological and innovation trajectories unfolding inside it. This contrasts with the mainstream of innovation indicators, which do not pay detailed attention to the composition of innovation activities.

[2] And even in this case there is debate about the definition of what comprises a 'chronic condition', as we will see below.

We also generate network indicators that help capture whether a mission-driven innovation policy has achieved the goal of creating new and sustained collaborations across disciplines, potentially developing consistent mission fields with a focus on pursuing a mission in the longer term. We are not aware of any mainstream R&I indicators that consider these structural implications of R&I policies.

We incorporate into our analysis social media data that measures social interest on the challenge being tackled in a mission. Again, this sets our approach apart from mainstream innovation indicators that do not consider public engagement or support for the goals of an R&I policy, and tend to use those data at a more disaggregated level in order to measure the reach of individual pieces of research in alt-metrics frameworks.

1.1.5 Stakeholder engagement summary

We presented our early thinking about this pilot in a Knowledge Stakeholder Workshop in Brussels in September 2019 involving representatives from the OECD, the European Commission, National Statistical Agencies and other policy and research audiences.

It was noted that there are important gaps in the evidence base in this area, and a very limited understanding of what design features of missions (for example in terms of the definition of its goal) are more conducive to success. Participants highlighted the need to select one level of analysis for our study, accepting that this will exclude some important modalities of R&I policy. We decided to focus our analysis on meso-challenges at a similar level of detail to those set out in Mazzucato (2018) or UK Government's Grand Challenges.

Another observation from participants was that the goals that R&I missions pursue can be very varied and likely to be captured in different data sources. This raises significant challenges for any attempt to develop a 'general purpose measurement framework' for R&I missions. Acknowledging this, we decided to focus our indicator development activities on R&I activities that are broadly shared across missions, acknowledging that fully capturing the impacts of particular missions will require the incorporation into this framework of bespoke components.

1.2 Relevance to RITO criteria

1.2.1 Relevant

The indicators developed in this pilot will help inform a new and important model for R&I policy where the evidence base is weak. The flexible system that we are building will allow policymakers to query the data to generate bespoke indicators adapted to their mission of interest, yielding highly relevant results.

1.2.2 Inclusive

Mission-driven R&I policy seeks to be more inclusive in the range of impacts that it pursues. By providing a stronger evidence base for these policies, this pilot will contribute to developing more inclusive R&I indicators. Further, the pilot will develop indicators that take into account the diversity of research trajectories in a domain and the expression of societal needs in social media. In doing this, it could help identify dominant avenues of development that exclude the interests, aspirations and needs of some groups, and misalignments between research trajectories and societal needs.

1.2.3 Timely

The pilot relies on open funding databases that are regularly updated its indicators will be timely.

1.2.4 Trusted

The pilot uses administrative data gathered by research funders in the UK and we do not have significant concerns about data quality: the data captures a population of interest for R&I policymakers. Our reliance on text data queried with keywords make the results transparent and easy to explain. Where we use less explainable methods, such as document embeddings to cluster projects, we rely on textual descriptions to facilitate interpretability. We make all our code available for review by others to enhance reproducibility.

1.2.5 Open

All our analysis uses open data sources and code.

1.3 Research/policy questions

Focusing on the UK grand challenge to "*Use data, Artificial Intelligence and innovation to transform the prevention, early diagnosis and treatment of chronic diseases by 2030*", these are the questions that we seek to address through the indicators that we are developing in this pilot:

What are the levels of activity and funding in this mission field?

How have the levels of activity evolved over time?

What is the disciplinary breakdown of the mission field?

How has the disciplinary breakdown of the mission field evolved over time?

What are the levels of interdisciplinarity in the mission field?

What is the distribution of outcomes in the mission field?

What actors are active in the mission field and what is their 'novelty'?

What is the diversity of technological trajectories in the mission field and how are they evolving over time?

2 Methodology

2.1 Data sources

The primary source of data that we have used in this pilot is UK Research and Innovation's Gateway to Research (GtR).[3]

GtR is an open, linked database with information about all research activities funded by Research Councils and Innovate UK (the UK's innovation agency), ranging from research grants in academia to innovation vouchers in industry. In January 2019 we queried GtR's open API and extracted all the information in the database. In total, this includes information about just under 90,000 projects and 38,700 unique organisations.

For each research activity, we potentially have information about its subject and its starting date, the organisations and individual researchers involved, the amounts of funding awarded, and the outputs, including publications, patents, spin-outs and technology outputs (such as software) to name a few.

Given our reliance on text descriptions to identify relevant projects, we focus our analysis on the 72,356 projects with an informative description in the corpus (this involves removing projects with missing and uninformative descriptions -eg. boilerplate text stating that a project description is not available for confidentiality reasons). It is also worth noting that GtR only covers research activity systematically since 2006. 98% of the projects in the data have a starting date in that year or later.

2.2 Methods

Data processing and enrichment

One area of interest for us is the level of disciplinary and industrial diversity in a mission field. However, projects in GtR are not classified into high-level disciplines. They are however tagged with a 'bottom-up' taxonomy of 607 research topics. We have built a research topic co-occurrence network that we analyse using community detection methods (more specifically, the Louvain algorithm) in order to identify sets of topics that tend to co-occur frequently, representing higher level, latent, disciplines.

The communities obtained through this analysis represent seven disciplines: 'arts and humanities','biological sciences', 'engineering and technology','environmental sciences', 'mathematics and computing', 'physics' and 'social sciences'.

In order to solve the problem of classifying projects into disciplines taking into account that interdisciplinary projects can belong to more than one discipline, we identify 'pure discipline' projects that only have research topics from one discipline community and combine them into a labelled dataset that we use to train a machine learning model that predicts project discipline labels based on the (pre-processed) text in their abstract.[4] We then use this model to predict the discipline mix for all projects in the data.[5]

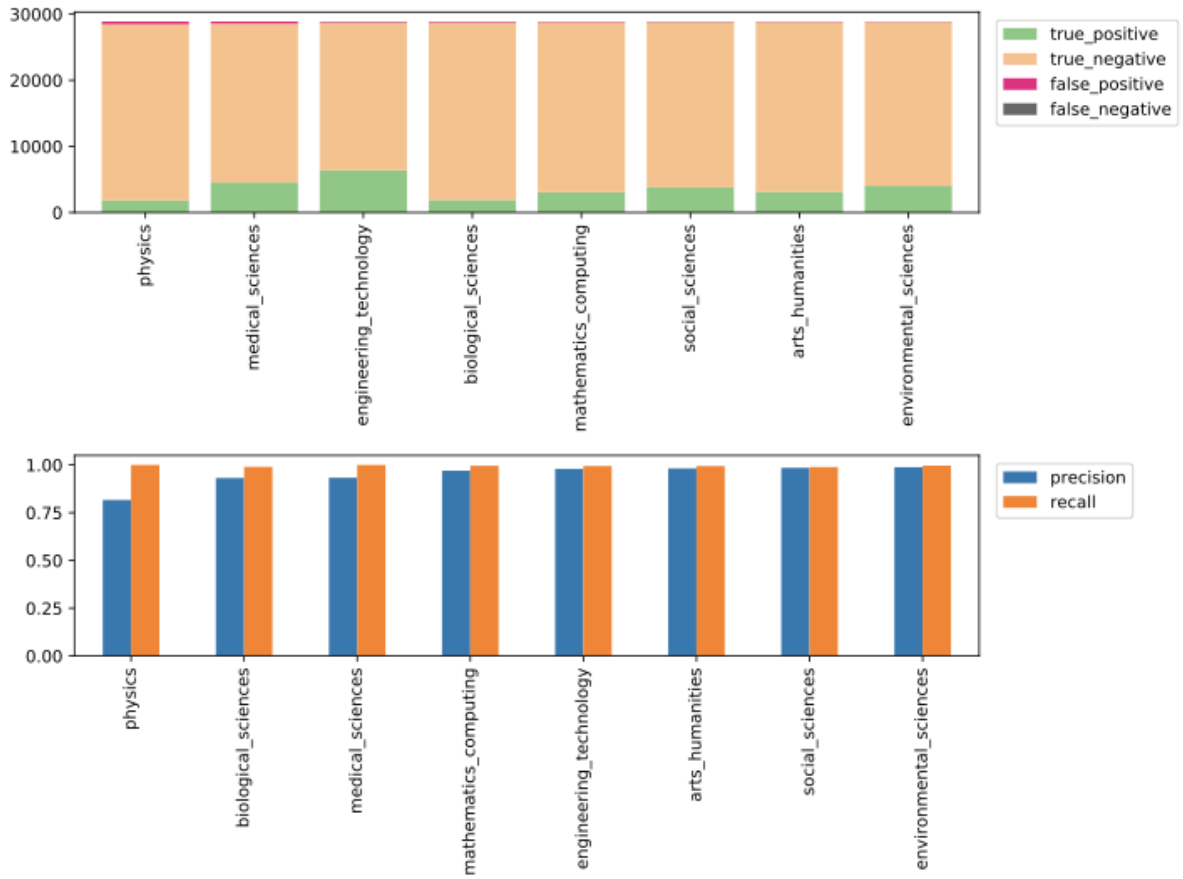
---

[3] https://gtr.ukri.org/

[4] 82% of the projects in the data have research topics that belong to a single discipline. Since the Medical Research Council does not use research topics to label its projects, we assume that all the projects it funds are in the Medical Sciences.

[5] We use grid search over a family of linear models and random forests with different levels of regularisation and leaf sizes, and three-fold cross validation. The best performing classifier is a logistic regression with L2 regularisation and balanced classes.

Figure 1: Stacked confusion matrix and precision and recall for discipline predictors



The two barcharts in the figure above show strong model performance for different disciplines (the only exception being physics, for the discipline classifier displays somewhat lower levels of precision and which we have not considered in the rest of our analysis).[6]

Mission semantics

In order to answer our research questions, we need to identify projects that are relevant for a given mission. We do this using the mission statement, in the case of this pilot, to "*use data, Artificial Intelligence and innovation to transform the prevention, early diagnosis and treatment of chronic diseases by 2030"*. This mission statement contains the following components:

A subject: Data, Artificial Intelligence and Innovation

A verb: to Transform

An object: Prevention, early diagnosis and treatment of chronic diseases

A timeline: by 2030.

[6] The number of Physics projects in the active mission field was in any case negligible.

A research project that was perfectly relevant for the mission would contain all of these elements in its abstract. It is however unlikely that we will find such perfect matches. For example, few research projects will focus on more than one dimension of the object. Projects will frequently specialise in a single chronic disease or group of chronic diseases. Few projects will specify the dates when they expect to generate applicable findings - that is a policy goal, not a research goal. We therefore distil the mission statement into the key components that we would expect to find in a research project abstract. They are the methodology they use (Artificial Intelligence) and the domain where they operate (Research on chronic conditions). From the point of view of the mission, the methodology is a tool, and the domain is a problem (health challenge) that the methodology seeks to 'solve' or at least alleviate. Note that we remove from the methodologies terms such as 'data' or 'innovation' that are quite generic and likely to appear in many irrelevant abstracts.

Having identified these mission components, we can define a potential mission field and an active mission field. The *'potential mission field*' comprises all projects that mention at least one mission component - in our case, the the solution (AI) *or* the challenge (chronic diseases). The '*active mission field'* comprises the projects that mention all mission components - in our case, the solution *and* the challenge: these are the most relevant projects from the point of view of the mission (see diagram below). One dimension of impact of a mission R&I policy is its ability to grow the active mission field at a faster rate that its mission components.

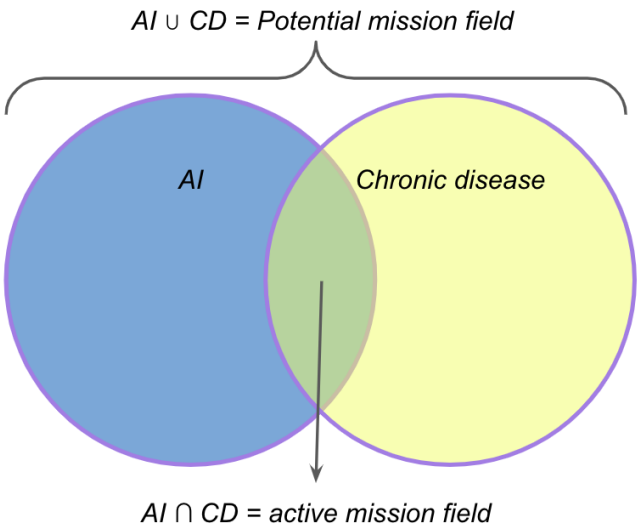


Figure 2: Defining and expanding the mission vocabulary

The first step in our analysis is to identify the list of keywords that capture different components of a mission - in our case, AI and chronic diseases. In the case of AI, we use the keyword 'artificial intelligence' as well as other analogous terms such as 'machine learning' and specific methodologies and techniques such as 'deep learning' or 'text mining'. In the case of chronic diseases, there is no consensus about their definition and the range of diseases and conditions that they include. Given this, we have opted for a crowdsourced list of conditions from wikipedia. This includes the following terms:

'Addiction, Aids, Alzheimer's, Atrial fibrillation, Autoimmune disease, Bipolar_disorder, Blindness, Blindness, Cancer, Cardiovascular disease, Cerebral palsy, Chronic condition, Chronic disease, Chronic hepatitis, Chronic pain, Deafness, Depression, Endometriosis, Epilepsy, Hiv, Huntington's, Hypertension, Lupus, Lymes disease, Parkinsons, Sclerosis, Sickle cell anemia'

We acknowledge that there is an element of arbitrariness in this definition. In future applications, the identification of the mission vocabulary should be undertaken in collaboration with the policymakers who defined the mission in order to ensure that the mission field analysis captures the relevant phenomenon.

In order to improve the recall of our queries (that is, our ability to capture relevant entities in the data), we perform a keyword expansion in our original list of terms. This expansion identifies other keywords that are similar to the terms in our original seed list based on their proximity in a word embedding space (Mikolov, Yih, & Zweig, 2013). This should increase the robustness of our results by helping us identify projects using synonyms for the words included in the initial search. This expansion results in a final list of 22 keywords related to AI, and 174 words related to chronic diseases, with which we query the data.[7]

The exhibit next page displays stage three random example descriptions of projects that contain both AI and chronic disease keywords. It suggest that even the relatively 'rough' implementation of our vocabulary selection and keyword expansion method yields relevant projects.

2.3 Documentation

All our code is available in this GitHub repo: https://github.com/Juan-Mateos/eurito_mission. We provide Jupyter Notebooks that describe the steps we have taken in the analysis, and our outputs.

[7] After an initial exploration of the keywords we remove a small number of highly generic ones such as 'fatal', 'syndrome' or 'software engineering' which would have introduced false positives into our analysis.

TITLE: Adaptive Automated Scientific Laboratory
====
ABSTRACT EXCERPT: Our proposal integrates the scientific method with 21st century automation technology, with the goal of making scientific discovery more efficient (cheaper, faster, better). A Robot Scientist is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experimentation. Our vision is that within 10 years many scientific discoveries will be made by teams of human and robot scientists, and that such collaborations between human and robot scientists will produce scientific knowledge more efficiently than either could alone. In this way the productivity of science will be increased, leading to societal benefits: better food security, better medicines, etc. The Physics Nobel Laureate Frank...

TITLE: New challenges in high-dimensional statistical inference
====
ABSTRACT EXCERPT: As a society, more and more of the activities that we take for granted rely on sophisticated technology, and are dependent on the fast and efficient handling of large quantities of data. Obvious examples include the use of internet search engines and mobile telephones. Similarly, recent advances in healthcare are partly due to improved, highly data-intensive scanning equipment in hospitals, and the development of new, effective drug treatments, which have been the result of extensive scientific study with data at its core.

Nevertheless, such advances can only be achieved through the development of appropriate statistical models and methods which enable practitioners to extract useful information from these vast quantities of data. In order to capture the complexity of the data generating...

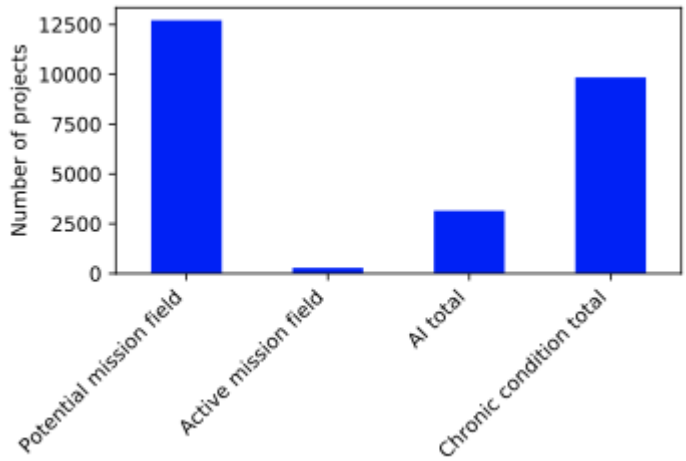
TITLE: Intelligent and Personalised Risk Stratification and Early Diagnosis of Lung Cancer
====
ABSTRACT EXCERPT: Lung cancer is the second most common cancer in both males and females, and has a very poor prognosis, causing >35,000 of cancer-related deaths each year (nearly 100 every day). This is due to the mostly very late-stage diagnosis of cancer: nearly 50% of all lung cancer cases are only diagnosed at very late Stage IV where no curative treatment exists. The annual cost of lung cancer to the UK economy is estimated to be around £2.4 billion, taking into account the cost of treatment and premature death, the cost to business of sick leave and of unpaid care by friends and family. It eclipses the cost of any other cancer, and continues to present a significant economic and healthcare burden. There is currently no national lung cancer screening programme in the UK, as current tests are...

3 Results

3.1 Findings

3.1.1 Mission field activity and evolution

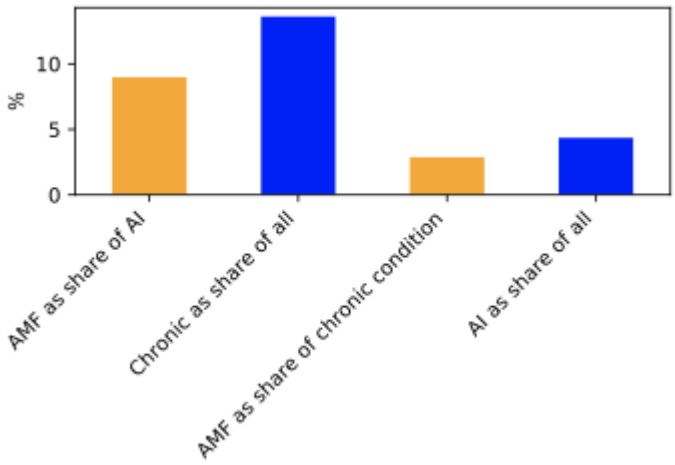
The figure below summarises the results of our classification: we have identified a potential mission field of 12,718 unique projects including any keywords related to either AI or chronic conditions. 3,152 projects mention AI keywords, and 9,849 projects mention chronic disease projects The active mission field (AMF) of projects overlapping both mission components is 283 (2.2% of the potential mission field).



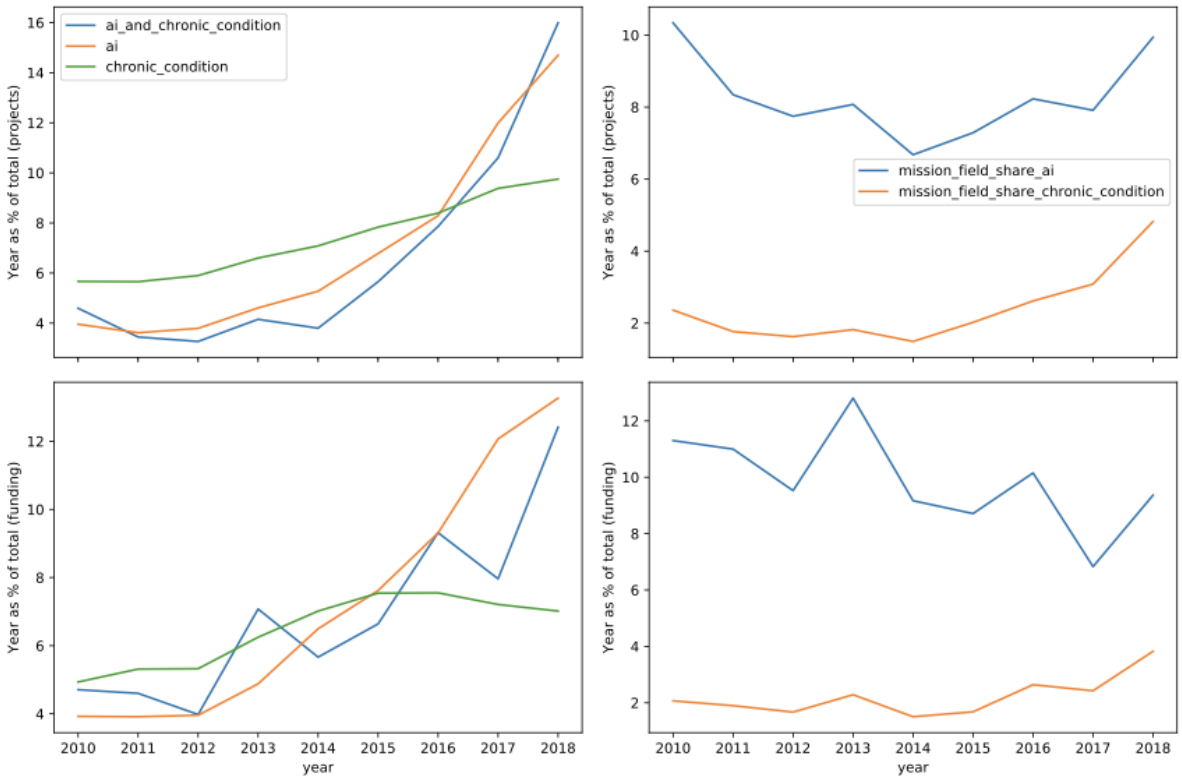
The figure below considers the representation of chronic disease projects in AI compared to the population of projects overall, and the representation of AI projects among the population of chronic disease projects compared to the population of projects overall. If AI methods were allocated randomly to application domains, we would expect the share of AI projects in chronic conditions to be the same as its share in the total. However, this is not the case. AI methods tend to be underrepresented in chronic condition research, and vice versa. We estimate that AI methods are 35% less likely to be used in the domain of chronic diseases than what we would expect given their overall distribution in the data.[8]

This result suggests that the motivation of the mission (to accelerate the adoption of AI methods in this health domain) is valid.

[8] Here, it is worth noting that some AI projects are likely to be basic and therefore not applied in any fields. It would be interesting to compare the representation of AI in chronic diseases with other application fields.



The figure below presents the evolution of activity around the mission field. The two line-charts in the left compare the share of all projects (first row) and funding (second row) in a category for every year. The two line-charts on the right consider the share of the active mission field in each mission component. We use 2-year moving averages in both cases to remove some of the volatility in the series.



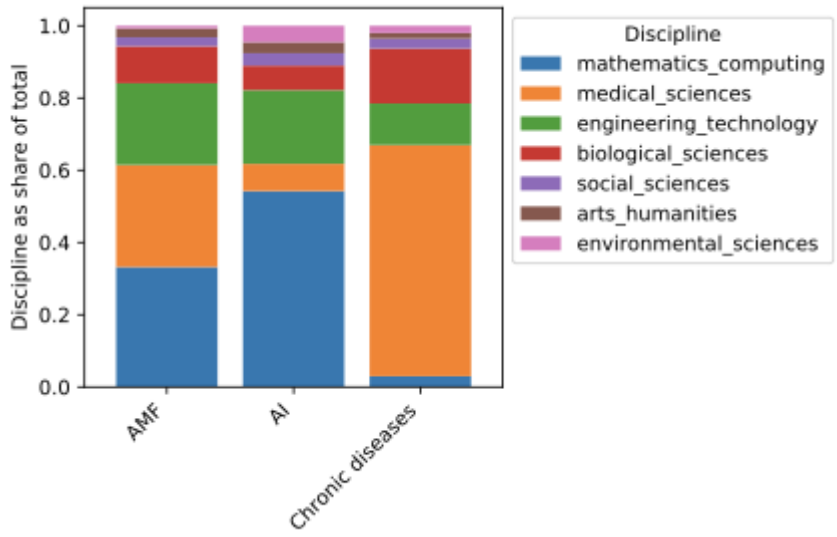
The first set of figures shows a steep increase in project and funding activity for AI and the active mission field in recent times (and particularly since 2014). By contrast, relative amounts of project activity and funding for chronic diseases has remained stable (or even declined) in recent times. We also see an upwards trend in the share of AI projects focused on chronic diseases, and in the share of chronic disease projects that use AI methods. This trend is less visible when we consider amounts of funding instead of number of projects, suggesting smaller, perhaps exploratory projects.

The analysis above suggests that the active mission field is developing as researchers start to apply AI methods in the field of chronic diseases - it is important to note that much of this activity precedes the announcement of the UK grand challenge in 2018, something to take into account when evaluating the impact of this challenge.

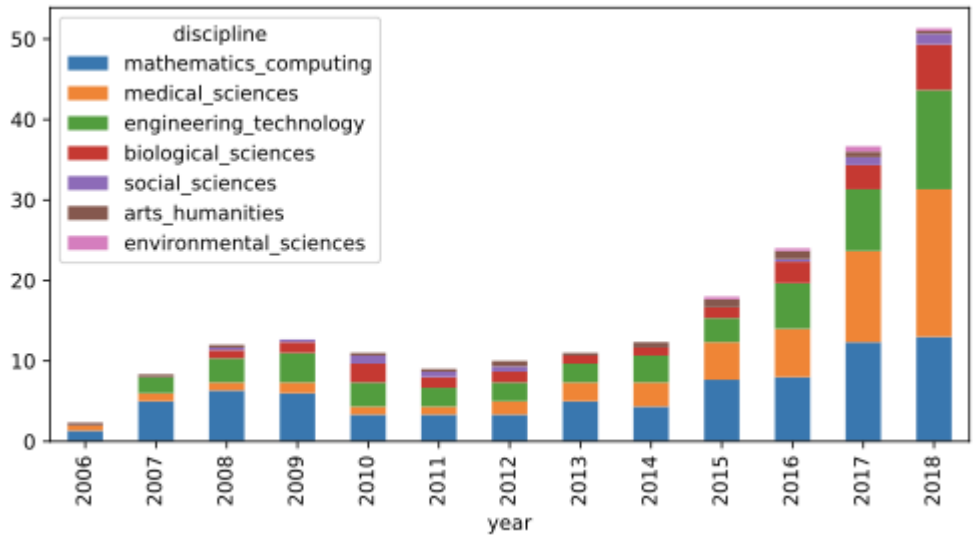
3.1.2 Mission field composition

What is the disciplinary participation in the active mission field?

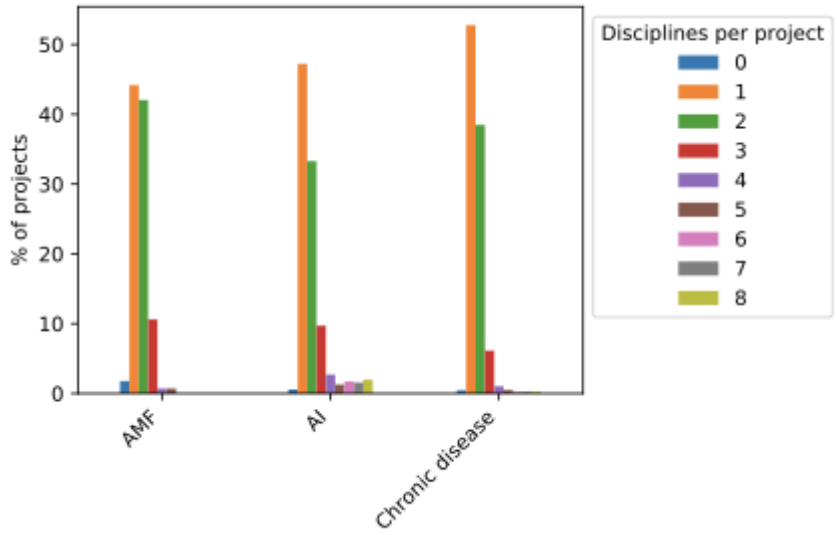
Below we display the discipline distribution of projects in the active mission field and the mission components after allocating each project to its discipline with the highest probability according to our machine learning classifier. The results show an almost even split between active mission field projects approaching the challenge from a Mathematics and Computing angle, and projects adopting a Medical Sciences perspective, followed by Engineering and Technology projects and projects from the biological sciences and biotechnology. By contrast, Social sciences and Arts and Humanities have very limited representation in the mission field, suggesting that social, organisational, policy and cultural dimensions of the use of AI to treat chronic diseases are not receiving much attention.



When we look at the evolution of the disciplinary mix in the active mission field, we see rapid growth in the number of medical science projects, suggesting that after an initial period where most applications of AI in the chronic disease area were dominated by computer scientists, AI methodologies are now starting to be deployed by medical practitioners.



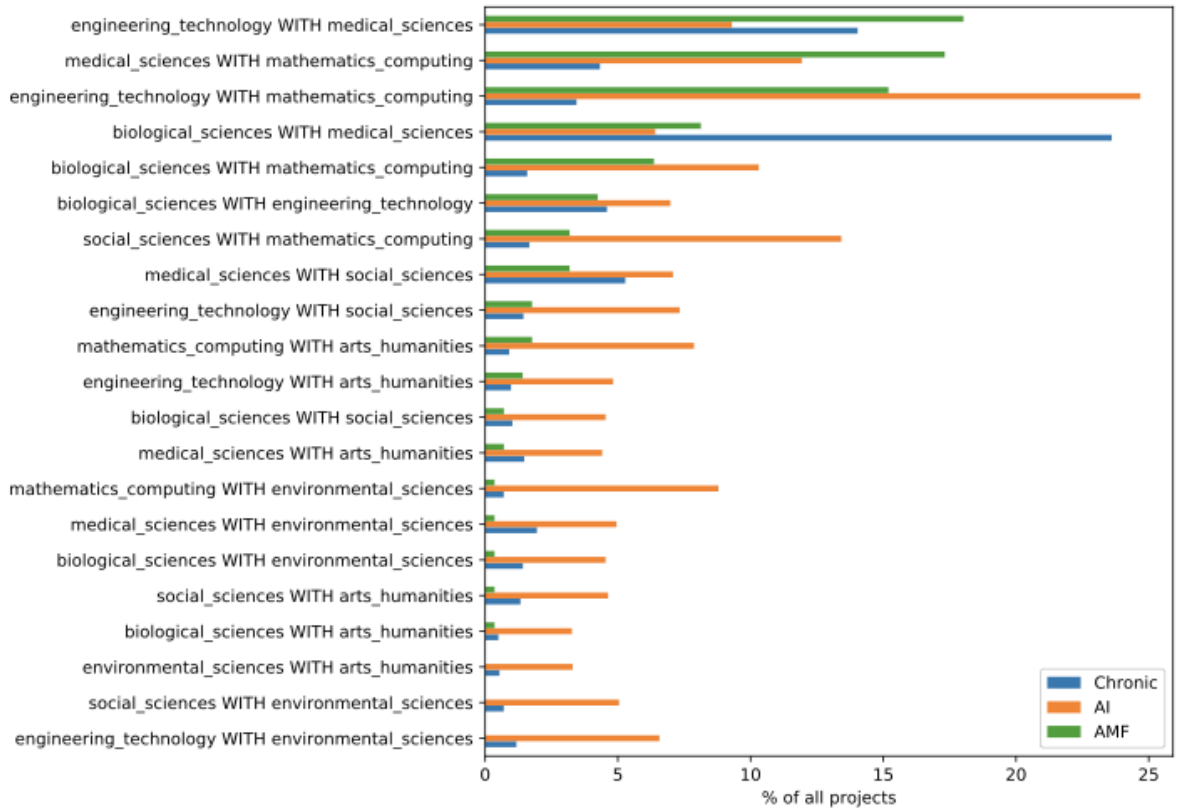
The analysis above has assumed that each project belongs to a single discipline. However, we can exploit the fact that our machine learning classifier generates a vector of discipline probabilities for each project in order to study discipline overlaps. The figure below shows the number of disciplines presents in projects in the active mission field and the mission components.[9] It shows that projects in the active mission field have a lower propensity to display a single discipline than projects in AI or Chronic Diseases, and a higher propensity to display projects with a presence of two and even three disciplines, consistent with the idea that activity in a mission field is more likely to involve broader disciplinary combinations.



Above we showed that projects in the active mission field tend to combine more disciplines. The figure below shows pairwise discipline combinations as a share of projects in the active mission field and its mission components. It shows that projects in the active mission field tend to combine Engineering and Technology with Medical Sciences, Medical Sciences with Mathematics and Computing, and Engineering and Technology with Mathematics and Computing. The figure also shows a stronger

[9] We set a threshold of 0.1 to determine whether a discipline is present in a project or not. The results are robust to changes in this threshold.

propensity towards disciplinary combinations in the AI field, also including non-negligible amounts of crossover with Social Sciences and Arts and Humanities. The main area of crossover in the Chronic Diseases mission component is between Medical Sciences and Biological Sciences.

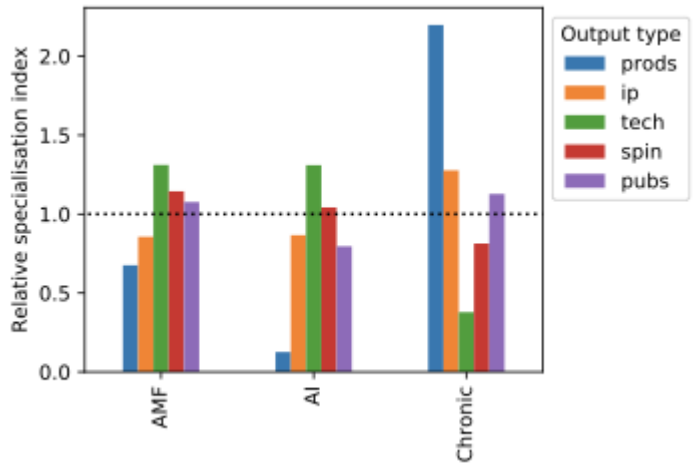


3.1.3 Mission field outputs

The criteria for mission selection we outlined above specify that a mission should be 'ambitious but feasible' - one way to explore this issue is by considering the ability of projects in the mission field to generate applied outputs. We explore this question with an analysis of output data available from GtR.

The figure below shows the 'revealed comparative advantage' of a field in generating different types of outputs. We produce this indicator by calculating the mean number of outputs per project in the different fields and normalising by the mean of all fields. A score above 1 indicates that, on average, a field tends to generate more outputs in a category than the others. We focus the analysis on projects started after 2014 to control for differences in field ages (as we showed above, active mission field and AI projects tend to be younger and therefore with less time to generate outputs than projects in the chronic disease field).

Our analysis suggests that the active mission field is technically feasible. Its projects tend to generate more technology (this dimensions captures software development activities in particular) and spin-outs than the average. Here, it is worth noting that the overall levels of output are low (for example only 10% of projects generate software outputs), and that the 'products' output category is predominantly used to identify biomedical products.

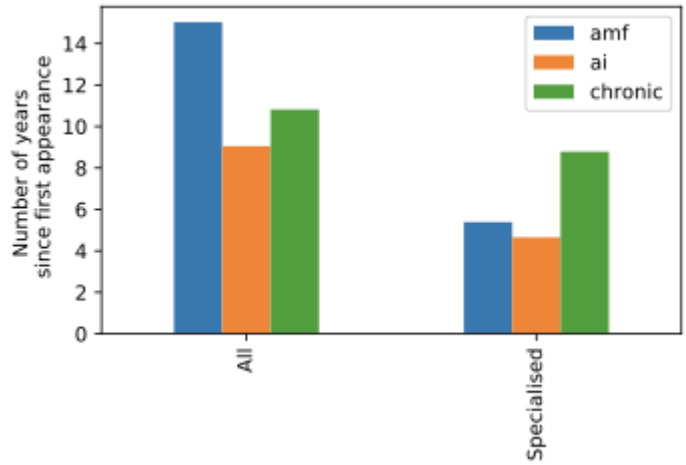


3.1.3 Mission field actors

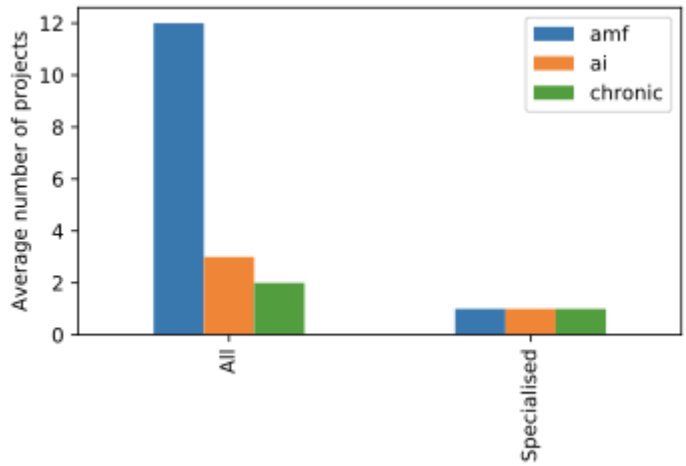
One of the goals of mission-driven policy is to lower barriers to entry for new actors in the innovation system. Here we develop indicators capturing whether actors in the fields we have been considering in our analysis (the active mission of field of AI *and* chronic diseases, and its mission components) are new entrants or seasoned veterans in the UK R&I system.

First we consider 'age' - here we want to compare the amount of time that organisations in different fields have spent in the R&I system as beneficiaries of research grants. To estimate this, we calculate the first year that an actor received a grant in our data and then average this value over all actors in a given field.

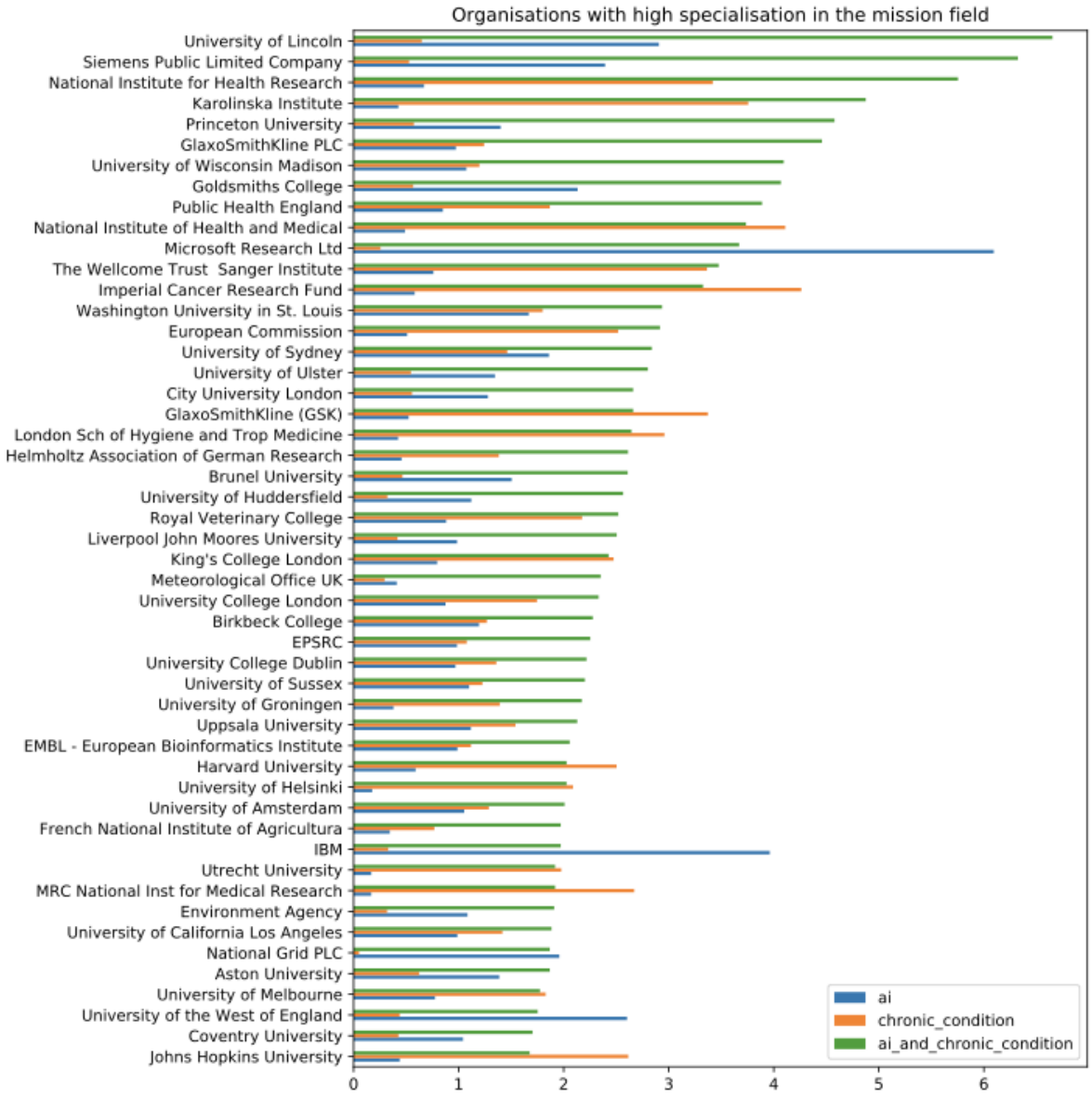
Contrary to expectation, we find that organisations participating in the active mission field tend to be older than those participating in AI and chronic diseases ('All' set of bars in the figure below). However, when we focus our analysis on organisations that participate exclusively in the mission field (thus excluding many large institutions that are active in it and many other fields), we find that the average time that an organisation has spent in the UK R&D system is lower for participants in the active mission field than is the case for organisations that participate exclusively in projects in the chronic disease field).



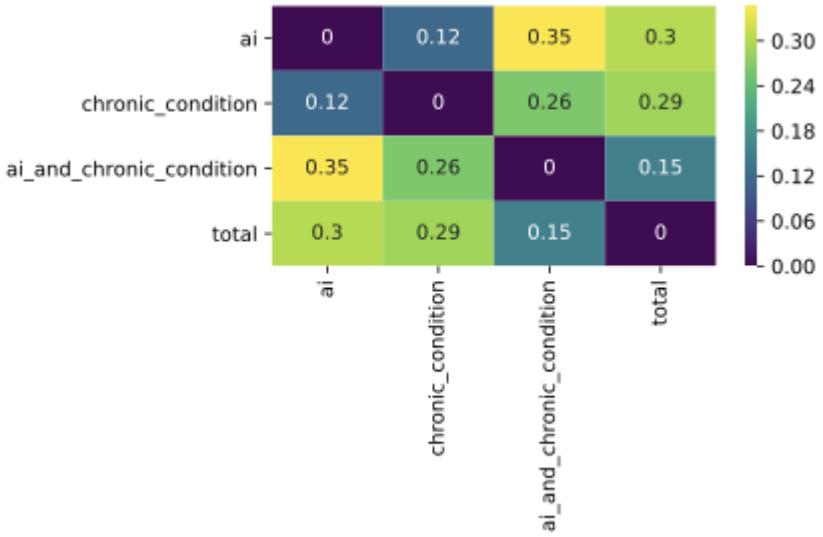
The figure below performs a similar exercise but focusing on the number of R&I projects that organisations in different fields have participated in, once again distinguishing between organisations that have participated in projects in a field *amongst other things,* and those that specialise exclusively in a field, and calculating medians instead of means to reduce the impact of outliers with thousands of projects (such as Russell Group Universities). Similarly to what we saw above, participants in the active mission field tend to be more active in the R&I system, while specialists are similar to organisations in other fields, with a median of project participation of one.



We conclude our analysis of mission actors by exploring some of the top organisations in the active mission field. Here we calculate relative advantage indices that normalise an organisation's share of activity in a field by its share of activity in all fields in order to capture whether it is specialised in a way that suggests a comparative knowledge advantage in a the field. One limitation of RAIs is that their values can be noisy for organisations with lower levels of activity so we focus on organisations with more activity. The figure below displays the 50 organisations with the highest specialisation in our active mission field, extracted from amongst the two hundred organisations with highest overall activity. We see that this list includes a varied range of universities, technology and biomedical companies, charities and public sector organisations. Interestingly, only six organisations in the top 50 of overall research activity appear in the top 50 for mission field activity, in line with the idea that a focus on missions might benefit less elite actors combining knowledge to address practical challenges.



Finally, we consider the link between an organisation's share of all activity in different mission fields, and in the total of publicly funded activity in the UK. The heatmap below presents bivariate rank correlations for those variables. Its results further reinforce the idea that activity around missions might create new opportunities for organisations outside the elite of the R&I system: the correlation between share of activity in a mission field and share of total activity is lower than for the mission components. Also we note with interest that the correlation between share of mission components and share of mission activity is stronger for AI than chronic conditions, suggesting that organisations with strong AI capabilities have been able to deploy them to address chronic diseases instead of the other way around, although this situation might be changing as the number of Medical Sciences projects applying AI methods increase.



3.1.3 Mission field trajectories

We finish the analysis in this pilot by considering the variety of research trajectories being followed in the active mission field. Here, we are interested in determining whether we are witnessing a parallel exploration of opportunities to apply AI to address chronic diseases, or to the contrary, if R&I activity is dominated by a few areas, perhaps ignoring some potentially fruitful application areas and patient groups. To explore this question, we cluster projects in the mission field based on their semantic similarity, and explore their composition and evolution.

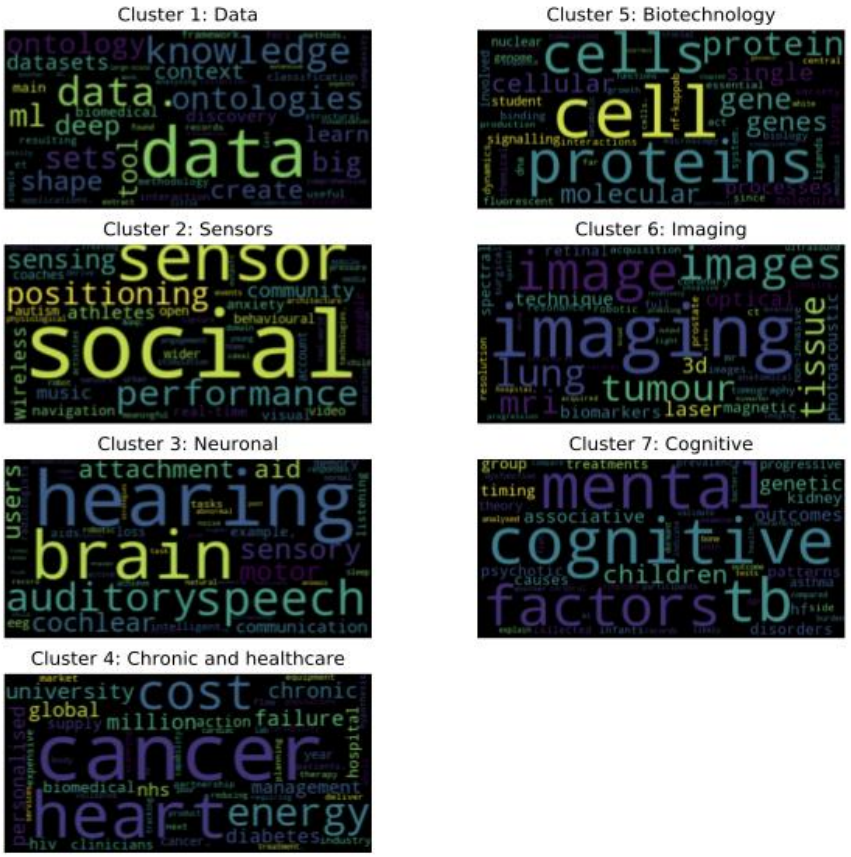
In order to cluster the projects, we embed them into a 200-dimensional vector space based on their semantic similarities using the Doc2Vec algorithm trained on the GtR corpus. Having done this, we identify clusters in the data using the Kmeans algorithm with seven clusters (this value for k optimises the mean silhouette score of the clustering solution after twenty runs).

Before we summarise some emerging findings from this analysis it is important to note that they are based on exploratory work that needs to be further validated and tested. In particular, we have chosen the Kmeans algorithm for practical reason and recognise the need to compare its results with other clustering algorithms based on different assumptions, as well as clustering algorithms such as DBSCAN that automatically select the number of clusters. Further, randomly initialised Kmeans clusters are not always robust in their classification of ‘edge cases’ between clusters.

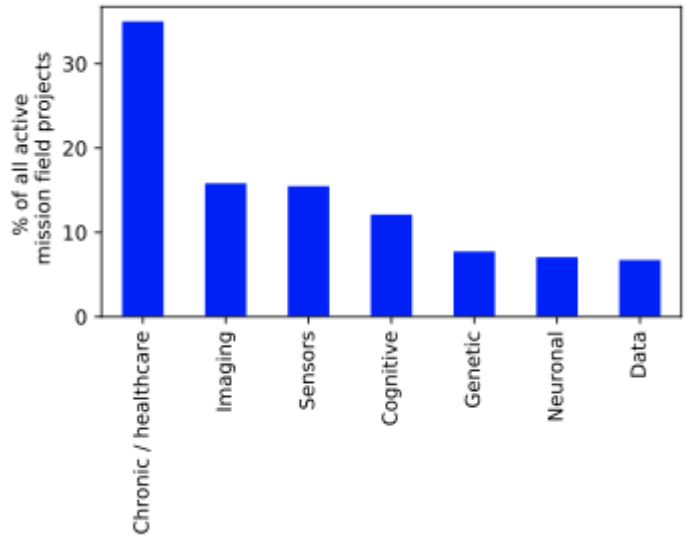
We plan to address this as a follow-up by running multiple iterations of an ensemble of clustering algorithms that we will then use to build a network connecting projects that tend to appear in the same clusters. We can then decompose this cluster into its constituent communities. One advantage of this approach is that it will help us understand the connections between clusters, something that is not possible with our current approach, where we allocate each observation to a single cluster.

Bearing in mind of all the above, our results seem intuitive. We display in the word-clouds below some salient words for the projects in every cluster we have identified. We identify one cluster centered on ‘big data’, another on the use of sensors and social data, one on the neuronal and sensorial systems, another about the ‘largest’ chronic diseases such as cancer, cardiovascular diseases and diabetes and also touching on healthcare applications, another focused on biotechnology and genetics, one using

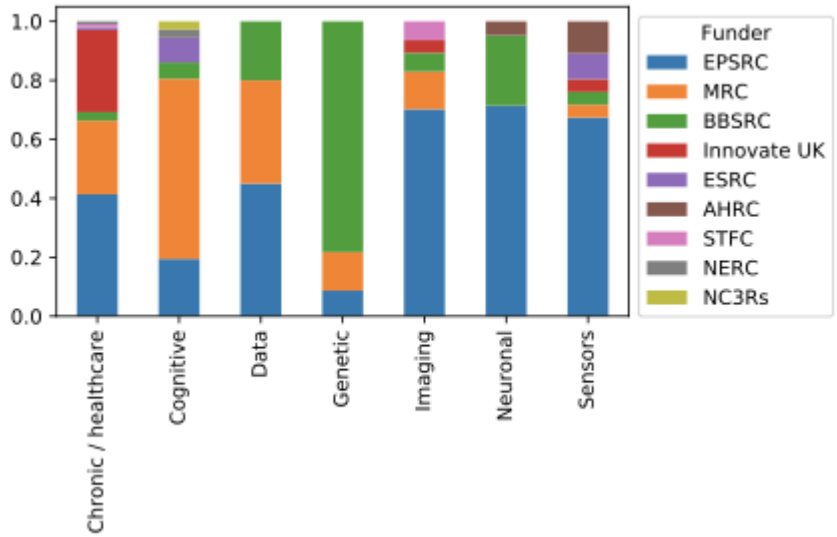
image data, and a final one that focuses on cognitive and mental issues. These clusters capture clearly defined application areas for AI in the treatment of chronic diseases.



As the figure below shows, although the 'Chronic diseases and healthcare' cluster of activity has the largest number of projects, the other application clusters also present a significant number of projects, suggesting that R&I researchers are exploring multiple avenues for deploying AI to address chronic diseases.

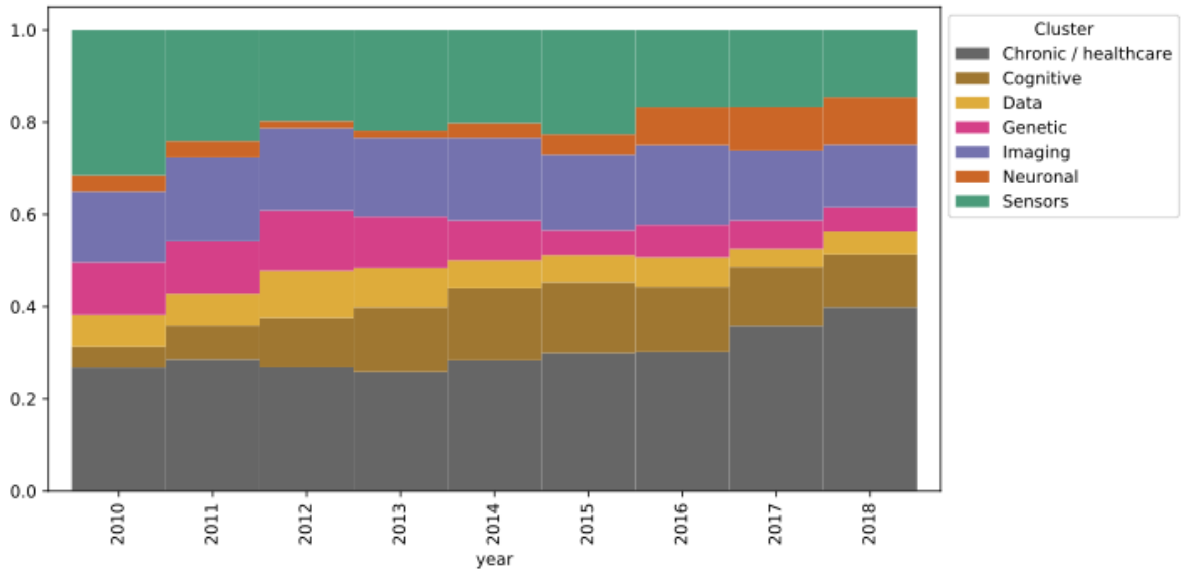


One potential reason for this might be the disciplinary specialisation of different funders, which is clearly visible in the figure below. Here, we see that different research councils focus their funding activities in different clusters: the MRC focuses most of the research in the cognitive area while the Biological and Biotechnology Research Council specialises on applications in Genetics and Biotechnology. The EPSRC dominates in device-centric areas such as Neuronal, Sensors and Imaging, while more general clusters such as 'Data' or 'Chronic and Healthcare' see participation from a wider range of funders.



We conclude by looking at the evolution of activity of different clusters over time. The stacked bar-chart below shows the share of all projects in the active mission field in different clusters based on the definitions above (we use moving averages to reduce the volatility in the series). We see that application areas such as Chronic/healthcare, Neuronal and Cognitive have gained relative importance in recent

years, while others such as Sensors or Genetic applications remain stable or lose relative importance (of course this in the context of rapid growth in the absolute levels of activity in the mission field that we showed before). This pattern is in line with the increasing importance of medical subjects in the mission field that we documented above.



4 Discussion and Conclusions

The exploratory analysis we have presented in this pilot illustrates the potential of analysing open funding databases to produce RITO indicators for mission-driven R&I policy.

In summary, we have shown that the UK mission to encourage the use of Artificial Intelligence to transform the prevention, early diagnosis and treatment of chronic diseases by 2030 is justified by the relative under-representation of projects related to chronic diseases amongst the population of projects that involve AI techniques. However, the number of projects in the active mission field have been growing rapidly in recent years, with increasing participation of Medical and Biological sciences projects. It will be important to remember this when it comes to evaluate the impact of the challenge.

Our analysis also reveals a high degree of interdisciplinary crossover in the mission field, consistent with the idea that different bodies of knowledge are coming together as AI is deployed to treat chronic diseases. When we look at research outputs, we find that projects in the active mission field are already producing applied outputs such as patents and spin-outs, suggesting that the mission field is technologically and commercially feasible.

We have also explored some basic demographic characteristics of participants in the mission field, finding that those specialising in it tend to be younger than organisations specialising in the chronic disease area. One interpretation of this is that the application of AI is bringing new 'blood' into that domain, again consistent with the goals of a mission (although not necessarily caused by it, as previously mentioned).

Finally, we have sought to identify different research trajectories in the mission field by clustering projects according to their semantic similarity. Our preliminary results suggest a diverse field with a variety of applications of AI to treat chronic conditions, ranging from big data projects to genetics, and sensors to cancer and cardiovascular. This reflects the diverse range of funders approaching the field from different disciplinary perspectives. Each of these trajectories could be analysed using the indicators we have developed in the pilot to get a more granular view of their composition, evolution and connectivity.

4.1 Validation and ongoing stakeholder engagement

In addition to further validation by colleagues at Nesta (including members of Nesta HealthLab with expertise on AI applications in health) and fellow members of the EURITO consortium, we will pursue several avenues to obtain stakeholder feedback about this analysis. This includes presenting and discussing emerging findings with UK organisations involved in the challenge (UKRI and BEIS), and submitting our analysis to relevant conferences in the fields of innovation studies and evolutionary economics.

4.2 Limitations

Our analysis is not with without limitations. First, and as discussed, our initial selection of chronic disease keywords has an element of arbitrariness - an important next step will be to validate and enrich this list with input from other stakeholders and domain experts, and in particular the research funders that organised the UK AI grand challenge. A better understanding of their scope of the domain and their specific objectives will help us hone our methodology and definitions further.

Our analysis is also constrained by its focus on open research databases that exclude non-academic research activities which could be relevant for a mission field, and do not generally include sufficiently

large project descriptions to deploy more sophisticated forms of semantic analysis than what we have done here. One option to address this would be to incorporate into our analysis text about project outputs, but this would be at the expense of timeliness, given publication and patent lags etc. We highlighted the potential for using social media data to capture public interest in a mission and its contents, and to analyse mission alignment with societal needs. We plan to do this in follow-on work.

The indicators that we have developed are relatively simple and therefore easy to communicate and explain, but this is at the cost of hiding some of the complexity of mission fields sitting in the intersection of domains. There are many opportunities to use network analysis to estimate structural characteristics of topic and organisational networks in mission fields and how they evolve in response to policy interventions, and will do this as we continue developing the pilot, also drawing on ideas and findings emerging from relevant work taking place elsewhere in EURITO (we highlight some connections with other pilots below).

This brings us to the question of impact, which as discussed in the introduction is currently not included in our analysis. Being able to attribute macro impacts such as future changes in patient outcomes in chronic conditions to a R&I policy intervention like the one we are studying here is a daunting challenge that will require the deployment of health-specific data sources and careful experimental design. Being able to monitor the journey of R&I outcomes from research into application could help with this - and data sources such as GtR or CORDIS, through Open AIRE, support this kind of analysis. At the same time, it will be critical to avoid losing sight of less direct but also important pathways to impact for the research funded as part of a mission, such as knowledge spillovers and expert labour flows. Geographical variation and clustering in the recipients of funding within the mission could help quantify these dimensions of impact, highlighting another interesting dimension of analysis (spatial) that remains underexplored in this pilot.

4.3 Considerations for scaling up

We already touched on the scalability of this pilot to other data sources and missions. As we said, significant components of the analysis we have undertaken here can be transferred to similar open research funding data sources such as CORDIS or the NIH's World Reporter as well as other R&I missions without significant costs in terms of data collection or infrastructure. Doing this would have the added advantage of creating comparators for indicator benchmarking and impact analysis. One potential problem with doing this will be the lack of standardisation in the language used to specify missions, which could make them difficult to compare. Perhaps the simple mission semantics that we have started developing in this pilot could help identify semantically similar mission types to be operationalised and evaluated using standard approaches.

On the dissemination and visualisation side, an important next step is to develop our offline and rigid system for querying innovation datasets for monitoring emerging technologies (in pilot 1) and mission fields (here) into Clio, an interactive system that users can use to query the data flexibly, adding and removing keywords in response to the results that are returned. Another important feature of Clio is that it will rank results in terms of relevance, helping the user manage the trade-off between precision and recall based on her own policy goals and preferences.

Interactive tools could simplify the selection of variables for analysis and exploration. As the preliminary results that we have presented make clear, the data we are working with have a many dimensions of interest (topic, time, funding, funder, participants, disciplines, connections with other organisations and topics) that could be combined into bespoke datasets and visualisations. Insights from

these explorations could be crucial for interpreting the data correctly, as our findings about the age of mission participants with different specialisation profiles demonstrate.

Another important consideration for scaling up is the collection of additional data from social media and sources related to mission impacts in order to cover new dimensions of a mission around social interest and alignment and impact that are currently absent from our analysis. We do not envisage significant data collection challenges for doing this, since we will primarily rely on well-documented and open APIs such as the CrossRef Event data, which we plan to use to understand social engagement with the outputs of mission-related research, and official data sources capturing impacts. Analysing these data in order to identify reliable and credible evidence of impacts will most likely be more challenging.

4.3.1 Complementarities with other pilots

There are strong complementarities between this pilot and Pilots 1 and 3 on emerging technology ecosystems and structural changes in technology. New technologies are often an important ingredient of R&I missions so the methodologies and tools being developed to monitor them in those pilots are relevant here. Reciprocally, some of the streams of analysis we have presented in this report, like the semantic analysis of R&D trajectories could be relevant in them.

Pilots X and Y on Advanced Funding Analytics and Knowledge Flows also approach open research data from different angles. The novel indicators developed in the first and the network analysis carried out in the second, as well as lessons from the analysis of H2020 data will be very relevant as we scale up this pilot.

There are also many connections between this pilot and Pilot Z on Inclusive Innovation. In particular, we are seeking to characterise inclusion in the purposes and impacts of innovation activity. It would be interesting to combine our analysis here with the study of the socio-demographic characteristics of innovators undertaken in that pilot in order to determine whether inclusiveness and diversity in team composition is linked to inclusiveness in the focus of research projects and their goals.

4.3.2 Tools and data sources

This pilot uses Python. All the analysis was performed in memory in a 16GB 2.4GhZ laptop and documented in JuPyteR notebooks. This repo [link] contains draft versions of the notebook still to be refactored and documented.

5 References

Acemoglu, D., & Restrepo, P. (2018). *Artificial Intelligence, Automation and Work*. National Bureau of Economic Research.

Aghion, P., David, P. A., & Foray, D. (2009). Science, technology and innovation for economic growth: linking policy research and practice in 'STIG Systems.' *Research Policy*, *38*(4), 681–693.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. Simon and Schuster.

Boschma, R. (2005). Proximity and Innovation: A Critical Assessment. *Regional Studies*, *39*(1), 61–74. https://doi.org/10.1080/0034340052000320887

Cantner, U., & Vannuccini, S. (2018). Elements of a Schumpeterian catalytic research and innovation policy. *Industrial and Corporate Change*, *27*(5), 833–850.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). *The Impact of Artificial Intelligence on Innovation*. National Bureau of Economic Research.

David, P. A. (1985). Clio and the Economics of QWERTY. *The American Economic Review*, *75*(2), 332–337.

Frenken, K. (2017). A Complexity-Theoretic Perspective on Innovation Policy. *Complexity, Governance & Networks, Complexity, Innovation and Policy*. https://doi.org/10.20377/cgn-41

Gustafsson, R., & Autio, E. (2011). A failure trichotomy in knowledge exploration and exploitation. *Research Policy*, *40*(6), 819–831.

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., … Morrison, A. (2018). The principle of relatedness. In *International Conference on Complex Systems* (pp. 451–457). Springer.

HM Government. (2018). *Industrial Strategy: Building a Britain fit for the future*. London: Department for Business, Innovation and Skills.

Jones, R., & Wilsdon, J. R. (2018). The Biomedical Bubble: Why UK research and innovation needs a greater diversity of priorities, politics, places and people.

Loder, J., & Nicholas, L. (2018). Confronting Dr Robot.

Mateos-Garcia, J. (2017). To Err is Algorithm: Algorithmic fallibility and economic organisation.

Mateos-Garcia, J. C. (2018). The Complex Economics of Artificial Intelligence. *Available at SSRN 3294552*.

Mazzucato, M. (2018a). Mission-oriented innovation policies: challenges and opportunities. *Industrial and Corporate Change*, *27*(5), 803–815.

Mazzucato, M. (2018b). Mission-oriented research & innovation in the European Union. *Brussels: European Commission*.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N13-1090

Mowery, D. C., Nelson, R. R., & Martin, B. R. (2010). Technology policy and global warming: Why new policy models are needed (or why putting new wine in old bottles won't work). *Research Policy*, *39*(8), 1011–1023.

Mulgan, G. (2017). *Big Mind: how collective intelligence can change our world*. Princeton University Press.

Nelson, R. R. (1977). *The moon and the ghetto*. New York: Norton.

Nelson, R. R. (2011). The Moon and the Ghetto revisited. *Science and Public Policy*, *38*(9), 681–690.

Ortega, J. L. (2018). Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics*, *116*(3), 2123–2138.

Restrepo, P., & Acemoglu, D. (2018). The Wrong Kind of AI?

Sarewitz, D. (2016). Saving science. *The New Atlantis*, *49*(Spring/Summer), 4–40.

Pilot 6: Advanced Research & Innovation Funding Analytics

Abstract

Since 1984 the European Commission has supported European Union research and innovation policy through its Framework Programmes (FP) and Horizon 2020 (H2020). The programmes have made billions of Euros available for public and private organisations to carry out research and development projects across a wide range of fields and applications, including science, technology, arts and the humanities. While high-level indicators have been introduced to assess the impact of the H2020 programme, it remains challenging to measure the outputs and outcomes from individual research projects. In this paper, we combine data from CORDIS, OpenAIRE and CrossRef to construct research assessment indicators at the project level. With this data, we are able to investigate the inputs to a project, such as the participant institutions, funding distributions, and project duration, and the outputs, including publication counts, publication numbers, and access types. In addition, we set out an approach to the considerations required for a healthy project-level research assessment framework, based on a review of literature that covers existing research performance indicators and assessment frameworks.

1 Introduction

1.1 Background and Context

The EU's H2020 programme has made €80bn available for research and innovation between 2014 and 2020. The funding programme is accompanied by a set of indicators for monitoring and evaluation that has been created in a bid to demonstrate the added value of research funded by the EU. The Horizon 2020 indicators report (Publications Office of the European Union 2015) contains a list of 23 compulsory key performance indicators (pages 11-15) and 14 cross-cutting issues (pages 16-19) which form the basis of the Horizon 2020 monitoring and evaluation framework. The indicators attempt to assess the high-level impact of H2020 research on European scientific and technological performance and research capacity, as well as any broader impact on the European economy and society.

While these indicators form a robust basis for an M&E framework they mostly consist of traditional indicators such as aggregates (e.g. total number of SMEs introducing market innovations) each evaluated independently of one another with no multivariate considerations or analysis of geographies, network properties of the interactions between actors, or the actual language content of outputs from a project. This aggregation masks the variety of outputs that can be created within research and innovation projects, making it harder to perform project-level evaluations and make decisions that take into account the diversity and value of different works across the portfolio.

In addition, the communication of existing indicators is limited to document based formats, which allow only for snapshots taken at a single point in time, and reduce the ability to interrogate statistics. While funding data about H2020 projects is made available through the Horizon Dashboard, generally the M&E indicators are communicated through documents published by the EU, such as the Interim Evaluation of Horizon 2020 (Directorate-General for Research and Innovation (European Commission) 2017). This is entirely appropriate for a high-level retrospective, but does not easily allow the reader to understand nuances and trade-offs between different possible dimensions of impact that exist within the research landscape.

For organisations seeking to engage in research and innovation evaluation, designing effective frameworks with appropriate indicators remains a challenge. In the United Kingdom, the Research Excellence Framework (REF) has come under criticism for threatening to undermine academic freedom, by tying research assessment to societal and economic impact (Martin 2011; Smith, Ward, and House 2011), for focusing on a narrow band of academic outputs, for being expensive (Dunleavy 2011), and for being open to 'gaming' by institutions (Gibney 2016). Excellence in Research Australia (ERA), the framework established by the Australian Research Council, aims to be "flexible and dynamic" and take into account the differences in outputs between academic domains, but has also been criticised by academics as being an onerous burden that distracts from research and other academic duties (Sardesai 2014).

There is clearly a gap for the creation of indicators that are complementary to the existing M&E framework that meet the RITO criteria (Relevant, Inclusive, Trusted, Timely and Open). With these, it will be possible to form a representative picture that can be used to make fair judgements about the success and impacts of research and innovation in the EU.

Designing evaluation indicators

The challenge of designing and implementing indicators and evaluation frameworks for research has become increasingly important as the digitisation of publications and proliferation of bibliographic databases enable the quantitative measurement of research outputs on a large scale. The new possibilities for metrics that are presented by these data sources must well thought out and responsible. While the systematised collection of equivalent data for the wider innovation sphere is relatively nascent, there exist commercially available databases about activity in the private sector, and governments are beginning to take advantage of their own administrative data sources. Whether the unit of analysis is a country, organisation, project or individual researcher, specific strategies for the creation and usage of metrics are required at each level.

Current research metrics and frameworks have been the subject of review and criticism from a variety of stakeholders in the research ecosystem, highlighting the need for continued work in this area and making it clear that there is no one size fits all solution. Several major limitations to current metrics have been analysed and identified, including their lack of correlation with peer review, the possibility of inadvertently creating counterproductive incentives, a narrow definition of qualifying research outputs and an inability to measure impact. (Lane 2010; Wilsdon et al. 2015) These flaws are often the result of a reliance on traditional bibliometric indicators, which have been designed for a specific academic discussion within the field of scientometrics rather than widespread real world deployment. In addition, the majority of data available is bibliographic in nature, limiting the types of measurable outputs, and data standards, such as the ORCID ID (a unique identifier for individual researchers) have not been taken up sufficiently or uniformly enough to always ensure accurate measures. Standardisations such as this one would enable a higher degree of reliability when attributing research outputs to researchers or institutions.

Despite this, or perhaps simply because metrics are now an embedded part of the research process, significant efforts have been made to create recommendations and guidelines for the next generation of evaluation frameworks. The Declaration on Research Assessment (DORA) also identifies the drawbacks listed previously, and sets out recommendations for assessing scientific output at the individual and institutional levels.(Cagan 2013) At the time of writing, the signatories of the declaration include 1190 organisations and 13731 individuals. The Leiden Manifesto is another similar such

document which arose during the 19th International Conference on Science and Technology Indicators. It cites the proliferation of metrics as "usually well intentioned, not always well informed, often ill applied".(Hicks et al. 2015) Both documents make suggestions for the responsible and useful application of metrics that are combined and summarised here, and found to be in keeping with the RITO criteria.

- Combine qualitative and quantitative - use metrics alongside expert judgement and traditional, trusted peer review methods for any in-depth assessment.
- Incorporate context - acknowledge and incorporate differences between research disciplines and geographical regions.
- Be transparent and accountable - be open about data collection processes and transparent on how metrics will be used, and allow those being assessed to be active maintainers of the information relating to them.
- Embrace variety - go beyond bibliometric indicators to capture the rich and varied outputs that can come from research.
- Ensure robustness and recognise limitations - identify feedback loops and incentives introduced by metrics. When using metrics to compare units of analysis, accept accuracy limitations and do not rely on false precision. For example, do not use digits after the first decimal to compare journals by their impact factor.
- Evolve with time - ensure that metrics are relevant to the current research system and consistent with the aims of all stakeholders by carrying out timely reviews.

Some funding bodies have already begun to echo these recommendations in updates of their evaluation frameworks. The Excellence in Research Australia system now aims to be a flexible and dynamic system that seeks to combine expert judgement with metrics, has introduced a standardised set of classification codes for disciplines, and attempts to assess both long and short term impacts of research. The Stern review of the REF has recommended that the next update of the framework should pay attention to interdisciplinary research, complement peer review with metrics and assess impact that is not only socioeconomic.

When considering the metrics that will exist within an evaluation framework, the nature of the stakeholders and the aims of the assessment should be considered critical. A 2013 report from RAND Europe offers ideas for 100 metrics that could be used, along with their caveats and limitations, as well as their suitability for different audiences and their relevance for four possible aims - analysis, accountability, advocacy and allocation, while highlighting that selecting any group of metrics to use will require compromises.(Guthrie et al. 2016, n.d.) Many of these metrics are variations of counts, which are open to biases and unsuitable simplification of research outputs, and so it is also important to examine how each of these can be improved. One such example shows the drawbacks of the commonly used h-index - Isaac Newton, Gregor Mendel and Peter Higgs would have h-indexes of 4, 1 and 9 respectively - as a response, a measure of citation impact that takes into account researcher age and co-authorship is suggested instead.(Belikov and Belikov 2015)

There is an opportunity to create a new generation of indicators that take into account these considerations, supported by existing official data sources and new open and web based data. For this pilot, this can be done by combining open EU research and innovation programme data with other open and web-based data sources, and new data science methods, to build a suite of indicators that fulfil the RITO criteria. The EU's CORDIS repository contains information about FP and H2020 funded projects, including details of the participant institutions, funding amounts, project descriptions and progress

reports. OpenAIRE, another EU data store has further information about projects and funding calls, as well as a database of publications that have resulted from the projects. This official data contain core details which can then be enriched with data such as MAK and CrossRef, which provide in-depth information on academics, publications, institutions and citations. Methods such as natural language processing (NLP) can be used to extract structured information from unstructured fields in the data, such as project titles, descriptions and reports.

1.1.1 Opportunity

This pilot attempts to develop more advanced metrics on European R&I funding by augmenting existing approaches with new variables such as skills diversity; generating new indicators that seek to capture intangible outputs using big data and machine learning techniques; and combining this indicator portfolio within an interpretable multivariate framework in which indicators are considered in unison and not isolation in order to understand the nuances and trade-offs present.

1.1.2 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

The availability of data used in this pilot and the fact that much of it is from the EU itself lends to the potential for this pilot to move from the periphery to the core of research and innovation policy. In addition, the growing movement to incorporate wider measures of impact into research assessment provides further reason to consider novel indicators. However, more robust data quality checks are needed before the findings can be transitioned fully to a decision making capacity.

1.1.3 Stakeholder engagement summary

This main stakeholder engagement for this pilot was carried out at the EURITO Knowledge Stakeholder Workshop which provided an opportunity to discuss questions and policy considerations around the advanced funding analytics pilot. These discussions formed the basis of understanding what kinds of information is often missing from research assessment within the EU. In addition, stakeholders had the opportunity to see a presentation from Frederique Bone (Sussex Policy Research Unit) which described a novel indicator for measuring various forms of diversity within a research collaboration.

1.2 Relevance to RITO Criteria

1.2.1 Relevance

The as Interim review of H2020 projects shows that many indicators are already being used to assess research, but they are traditional in nature and are exhibited in a report style format.

1.2.2 Inclusive

The aim of this pilot is to create indicators that capture the full range of research and innovation outputs, accounting for differences that can occur between subjects and regions.

1.2.3 Trusted

We aim to use a combination of trusted official alongside any external data.

1.2.4 Timely

The data that are explored in this pilot are updated by automated means, with some human review, ensuring that they are regularly up to date.

1.2.5 Open

The code and results for this pilot are openly available, and the data that has been used can be freely obtained.

1.3 Research and Policy Questions

Using the features of the CORDIS dataset and publication outputs collected from OpenAIRE, we propose to investigate the factors influencing research outputs, such as levels of funding, numbers of participating institutions, the types of participating institutions, and subject area. As discussed during the stakeholder engagement workshops, there are many heterogeneities within the funding landscape, such as competing priorities of among stakeholders. We therefore aim to discuss the relevance of variables to different actors and objectives.

A further policy question relates to whether the research that was funded was performed, and if the language used in outputs differs from that used in the original proposal. By collecting the text for project proposals and of project outputs, we will use Natural Language Processing techniques such as topic modelling to assess the similarity of proposals and outcomes whilst controlling for different language likely to be used in the two contexts. Additionally, we aim to use project keywords to distinguish between work that is novel and that which is highly cited, in order to demonstrate another dimension of output that goes beyond relying on citations or counting of outputs.

2 Methodology

2.1 Data sources

The data sources used in this pilot are from CORDIS, OpenAIRE, Crossref and MAK.

CORDIS is the European Commission's source of information about projects funded by the EU's framework programmes and H2020. It is a repository of data that includes funding amounts, descriptions, progress reports, funding levels and dates for the projects. It also includes information on each participant involved in a project, such as their name, address, institution type, and location. The data are made available in dumps that can be downloaded from data.europa.eu, and covers FP1 to H2020. For this pilot we examined the data around FP7 and H2020, spanning the period from 2007 to 2020 and covering 46,292 projects.

OpenAIRE is an organisation that seeks to further openness and transparency in science, including linking datasets relating to scientific research and its outputs. The website for their 'Explore' tool claims to include 25 million publications and other data from 14,000 content providers across 18 funders, including the EU. This website was scraped for publications that are linked to FP7 and H2020 projects. Each OpenAIRE publication entity contains title, publication identifiers, publication type and access options. 300,000 publication entities were retrieved. The funding data in OpenAIRE are obtained directly from the EU, while data on research results, such as publications are taken from content providers such as publishing houses. Researchers are also able to log in to the platform and link their research results, such as publications and datasets, to each other and to their own profile. Due to its mission, OpenAIRE content guidelines favour research that has been published openly, and therefore may express biases towards disciplines that already have an established ecosystem for open access publishing, such as those that use open repositories like arXiv.

Crossref is a non-profit that aims to make it easy to search for and cite literature. It provides an API endpoints to retrieve data about publications including metadata and events. The metadata includes the information required to cite a publication, as well as the number of times it has been referenced in other works, and the names of the authors.

Microsoft Academic Knowledge (MAK) is a linked data source that contains information on papers, authors, institutional affiliations, fields of study, conferences and journals. It is accessible through an API. We have used it specifically to collect information on papers, to fill gaps where Crossref does not contain information.

2.2 Methods

Data Linking

In order to enrich publication entities obtained from OpenAIRE with Crossref metadata, we required a way in which to link them. We chose to rely on the document object identifier (DOI) to link entities, as it provides a unique and reliable identity for each publication. However, OpenAIRE does not provide DOIs for all publications. In the absence of a DOI, many publications had PubMed identifiers (PMID), while others had nothing at all. Both of these issues were solved with a separate solution.

PubMed Central (PMC) provides an 'ID Converter API', which is able to retrieve the the corresponding DOI from a PMID (and vice versa). This API was queried for all publications with a PMID in the OpenAIRE dataset.

For documents with no unique identifier, we relied on the publication titles. The 'works' endpoint for the Crossref API exposes a search function, that returns the most likely matching publications based on a title query. For each OpenAIRE publication with no identifier, we queried this endpoint with the work's title to obtain a shortlist of matching titles. We then calculated the Levenshtein similarity between the actual title and those of the top 5 results, and returned the DOI for the work with the highest score. To reduce the number of false positives, we applied a minimum similarity threshold of 0.9. Using this methodology with Crossref's metadata API, it was possible to gather metadata to enrich 215,509 of the publications collected from OpenAIRE.

Every publication in OpenAIRE also contains funding metadata, including a European Commission project ID. Using the final publication data and linking by project ID, we were able to find papers for 21,176 projects.

Discipline Assignment

As part of the Mission Oriented R&I EURITO pilot, a machine learning classification model was created, to assign academic disciplines to research projects based on their descriptions. This classifier was also applied to the FP7 and H2020 projects in this pilot, in order to aggregate statistics by discipline. A model was created based on the methodology and data that can be found in Section 2.2 of Mission-oriented R&I: Pilots Research Results.

Minor modifications were made to the model: the algorithm chosen was a logistic regression train on the best 40th percentile of tf-idf features as measured by calculating chi2 between the features and target variables. The confusion matrix and classification scores for the classifier can be seen below.

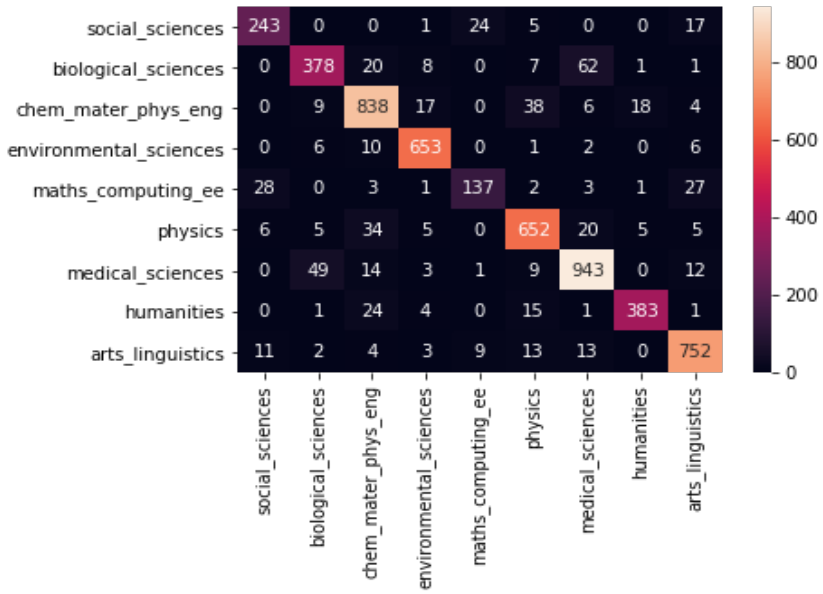


Figure 1.1: Confusion matrix for subject classification model.

We can see that most subjects in the test data were classified correctly, though there is some mislabelling. However, this is generally confined to subject areas where we might expect an overlap anyway, such as the mislabelling of medical science as biological, or physics (astronomical and particle) with chemistry, materials, other physics and engineering.

	precision	recall	f1-score	support
arts_linguistics	0.84	0.84	0.84	290
biological_sciences	0.84	0.79	0.82	477
chem_mater_phys_eng	0.88	0.90	0.89	930
environmental_sciences	0.94	0.96	0.95	678
humanities	0.80	0.68	0.73	202
maths_computing_ee	0.88	0.89	0.88	732
medical_sciences	0.90	0.91	0.91	1031
physics	0.94	0.89	0.92	429
social_sciences	0.91	0.93	0.92	807
micro avg	0.89	0.89	0.89	5576
macro avg	0.88	0.87	0.87	5576
weighted avg	0.89	0.89	0.89	5576

Table 1.1 Classification scoring for subject classification model.

The classification scoring shows weighted average precision, recall and f1-scores of 89%, though this varies across subjects. Humanities and biological sciences have the lowest recall, though further investigation into why this was the case has not been carried out.

Limitations

A primary limitation in this analysis is likely to be our ability to link and disambiguate between multiple data sources - e.g. identifying which “John Smith” is a chemist and which “John Smith” is a biologist and which of the two is the author of a particular paper. It is impossible to do this perfectly therefore care needs to be taken to minimise noise introduced at this stage of the pipeline. Part of this is mitigated by utilising existing, sophisticated solutions such as MAK, Crossref, OpenAIRE developer API which implement parts of this process, whilst augmenting these with additional layers of verification and analysis.

There is also a lack of counterfactual/control data - in an ideal world we would have data from a cohort of unfunded projects to evaluate future outcomes/outputs of the unsuccessful institutions and researchers. We currently have no way acquiring this data (though it exists within CORDA, the internally facing EU database behind CORDIS), therefore we need to be careful about the inferences we draw with one possibility being to compare to the average researcher/institution within a certain context (e.g. discipline/geography).

2.3 Documentation

All our code is available in this GitHub repo. We provide Jupyter Notebooks that describe the steps we have taken in the analysis, and our outputs.

3 Results

3.1 Outputs

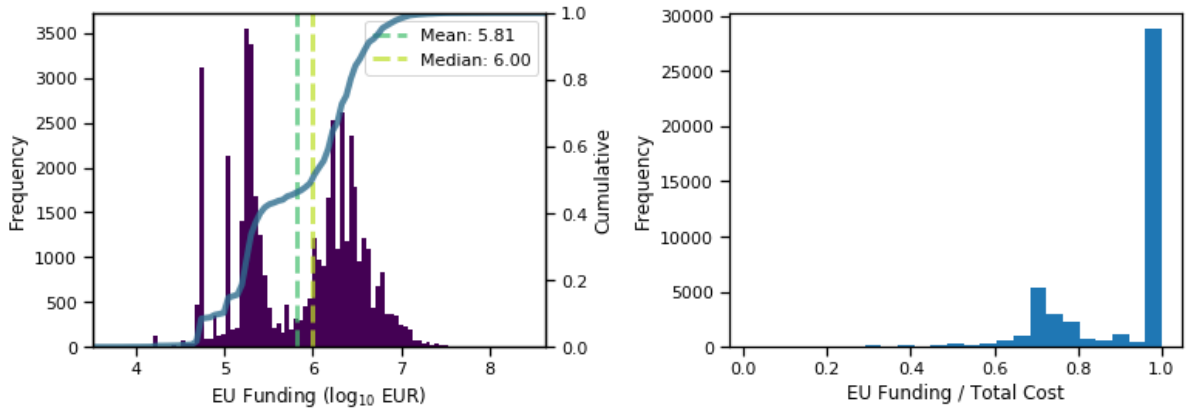


Figure 2.1: Distribution of EC funding in FP7 and H2020 projects

The distribution of EC funding across all projects shows a spiky profile with an underlying bimodal characteristic. This highlights that there are two broad funding categories; those in the order of €100k and those in the order of €1m. Within this distribution there are particularly high frequencies of individual grants centred on €50k, €100k, €250k, €2m, €2.5m, €3m and €5m amounts. Looking at the ratio between EC project contributions and total project costs, we can see that in the vast majority of cases, 100% of funds come from the EC. Other projects receive between 0 and 85% of their funds from other sources, with the 20 to 25% being the most common fraction contributed from third parties.

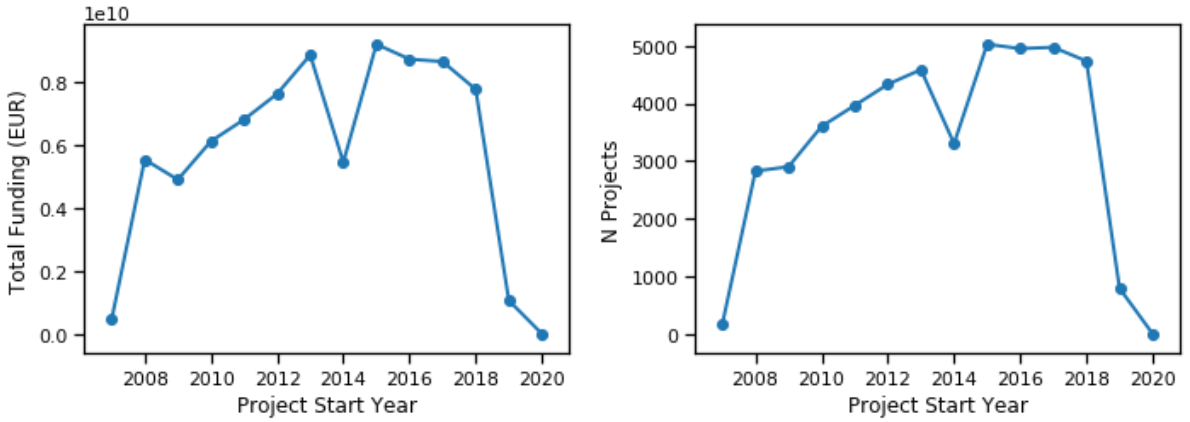


Figure 2.2: Total spend and number of projects in FP7 and H2020

The total funding awarded to projects by their starting year shows that generally between €4bn and €12bn have been awarded annually. The lower figure in 2007 may indicate overlap with the previous funding programme. There is also a dip in 2014, which is when the H2020 programme began, so this is perhaps indicative of the FP7 projects finishing and the new wave of projects not having yet started. Finally, we are presented with a decline in the amount in 2019 and 2020. This is likely due to the H2020 programme ending in 2020, and therefore fewer, shorter projects are due to start in those last two years.

The number of projects starting each year closely mirrors the trend in the amount spend, suggesting that the average spend per project has remained roughly the same over the years.

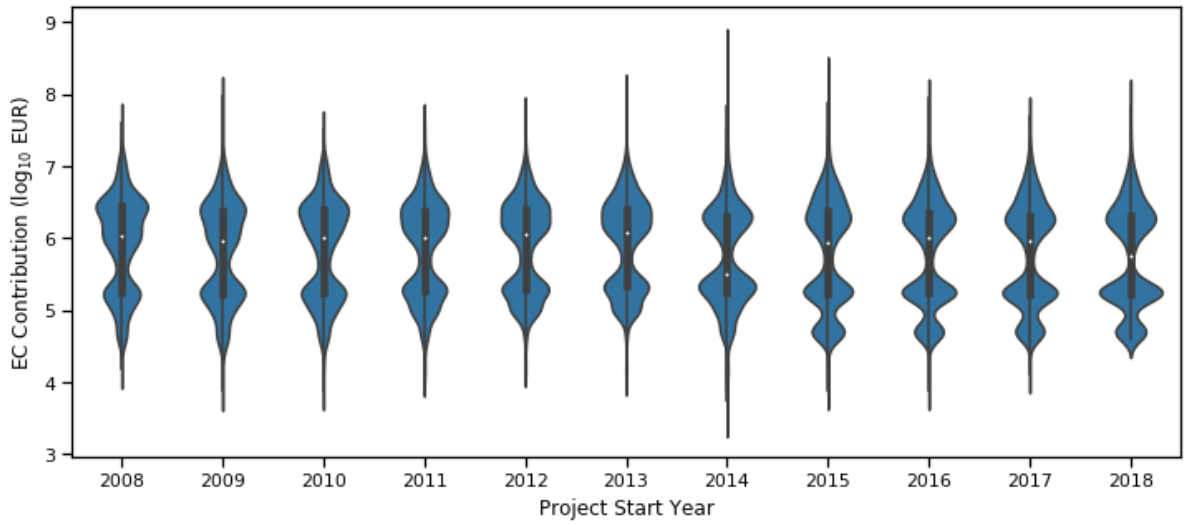


Figure 2.3: Distributions of EC project funding from 2008 to 2018

A violin plot can show us the change in funding allocations over the start year of the projects. We can see between 2008 and 2014 the more or less bimodal distribution seen earlier, with peaks centred around €250k and €2.5m. From 2015 onwards, we can observe the appearance of a third peak below €100k. This is perhaps a new H2020 initiative to fund early career researchers.

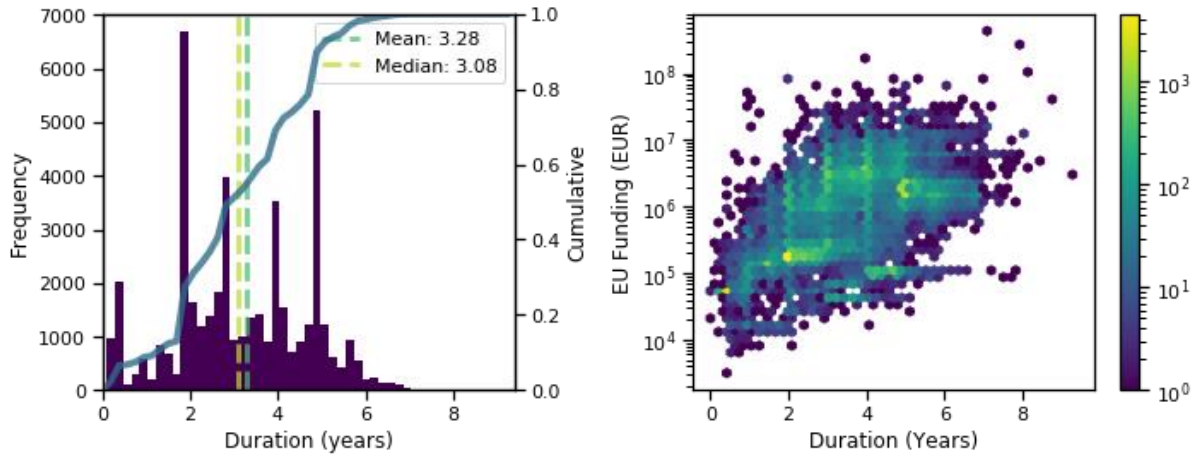


Figure 2.4: Durations of FP7 and H2020 projects

Project durations cover a span from a few months to several years, with major peaks at 2, 3, 4 and 5 years, and a mean length of 3.28 years. Typically longer projects are awarded more funding. For example, we can see from the peaks at A project designed to run for 5 years might receive up to 10 times more funding than one set to last for just 2 years.

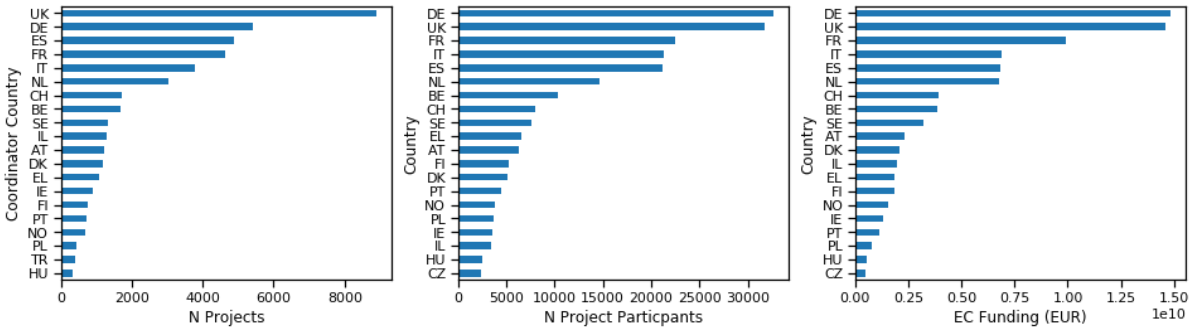


Fig 2.5: Number of projects, participants, and amount of funding by country

The UK is the coordinating country of the most awards, accounting for almost 9,000 projects. Only Germany, Spain, France, Italy and the Netherlands also coordinate more than 2,000 projects each. However, when this is broken down by the number of times an institution from a country has participated in a project, we can see that Germany just overtakes the UK, with both having over 30,000 institutional involvements in a project. The top 6 countries are again the same. The same 6 countries occupy the top spots when it comes to the total amounts of funding received from the EC. Again, Germany and the UK are very close, both having gained just over €14bn, and France is the only other country to have been awarded a total greater than €10bn.

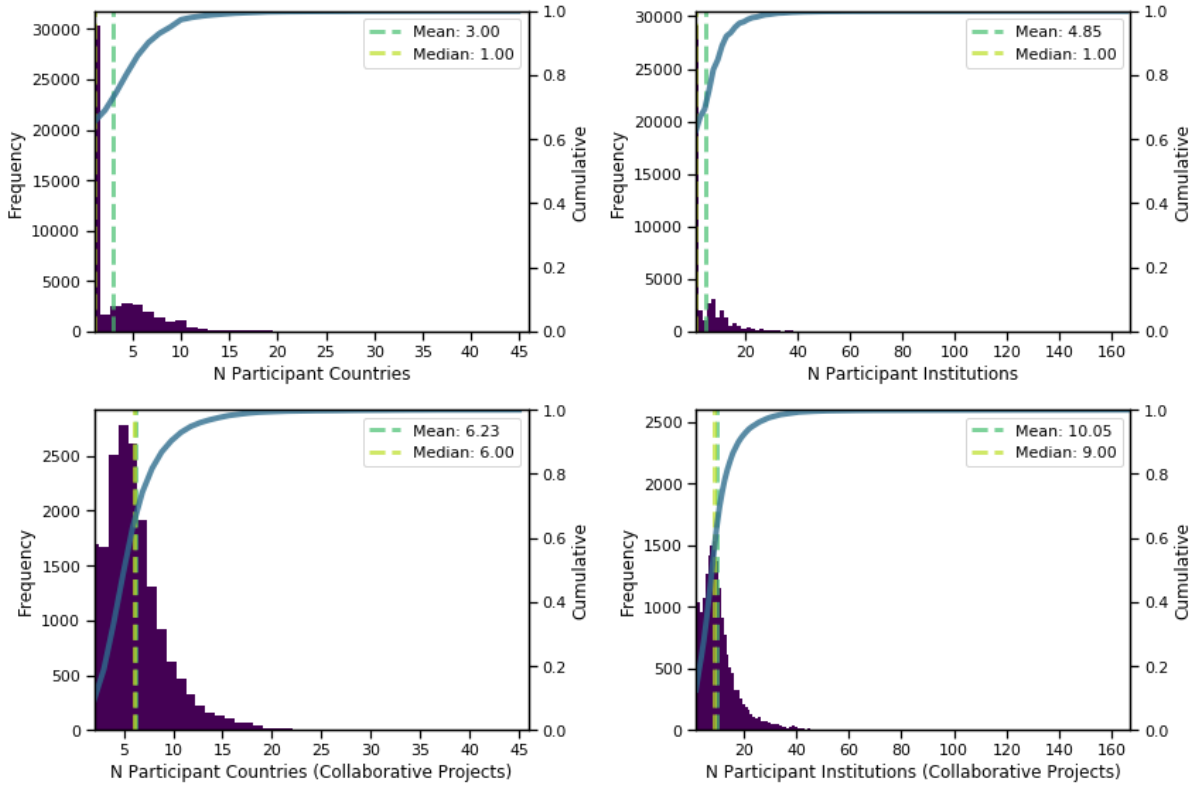


Figure 2.6: Distributions of numbers countries and participants involved in FP7 an H2020 projects.

The distributions of the numbers of countries and institutions involved in each project shows a spike at 1, suggesting that most projects (around 60%) are not collaborative between institutions. We can see, however that in both distributions, there is a secondary peak at higher numbers, and that the distribution of countries involved in a single project goes up to ~45 and the distribution of participants involved

spans to over 160, showing that some projects are the product of large numbers of institutions. If we look at collaborative projects only (those with >1 participant country or institution) we can see the distribution of numbers involved in multi-party projects. At least 50% of projects involve 6 countries or 9 institutions. This shows that for collaborative projects, the EU tends to favour funding multilateral projects, rather than simple bilateral collaborations.

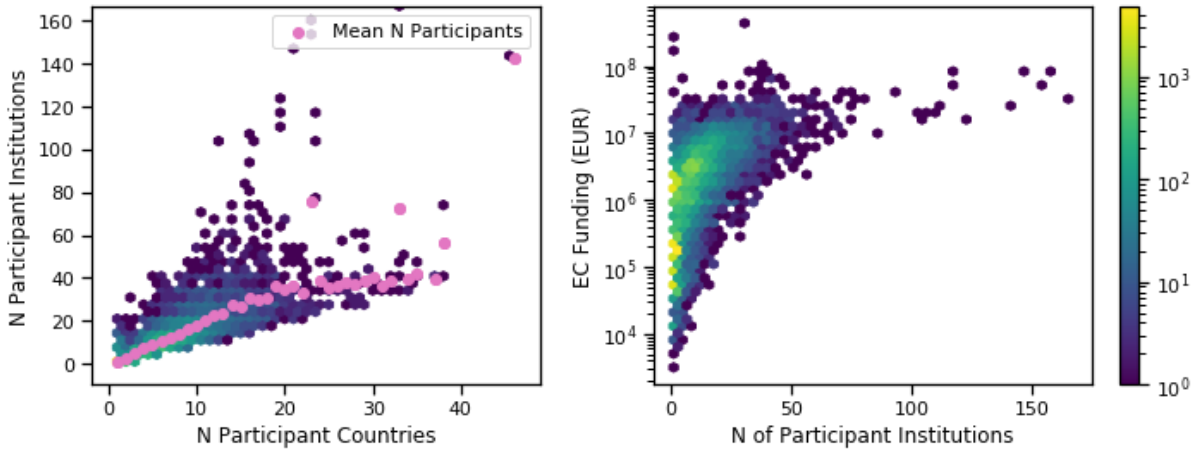


Figure 2.7: Numbers of participant countries and institutions per project, and corresponding levels of funding.

There is a roughly linear relationship between the number of countries participating and the number of institutions involved in a project. The modal value is of a course a single country and organisation. We can see that the maximum number of institutions involved in a project can often exceed the number of countries, meaning that there must be multiple institutions from a single nationality involved in a proportion of projects. Plotting the mean number of organisations on a project against the number of countries involved, we can see that the number of participants is roughly equal to twice the number of countries up until 20 nations are involved. After this, we see a tapering off of this trend.

The funding distribution across projects with small numbers of participants involved resembles that of the overall funding distribution seen in Fig. 1, and has projects ranging across the whole spectrum of funding amounts. There are a significant number of projects carried out between 0 - 25 institutions that receive multiple millions of Euros. As the number of participants increases, the project eligibility for smaller grants appears to disappear. For example, only a small number of projects with more than 25 participants received a grant of less than €1m.

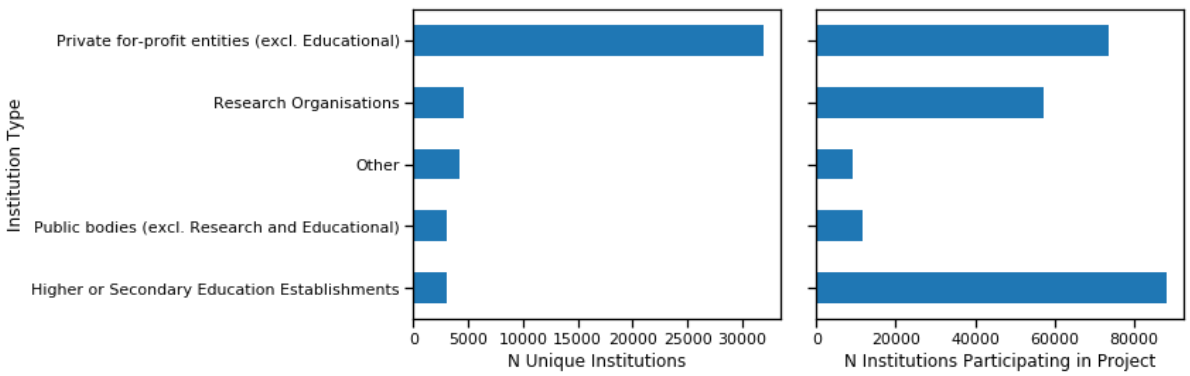


Figure 2.8: Counts of institutions in FP7 and H2020 projects by type.

Each organisation involved in a project listed in CORDIS, has an associated activity type, which corresponds to the type of institution that it represents. The categories are *Private for-Profit (Excluding Education Establishments)* (PRC), *Research Organisations* (REC), *Other* (OTH), *Public Bodies (Excluding Research Organisations and Education Establishments)* (PUB), and *Higher or Secondary Education Establishments* (HES). By counting the numbers of unique institutions in the dataset, we can see that there are over 30,000 private for-profit organisations, and less than 5,000 in each of the other categories. However, when we look at the number of instances of each institution type being involved in a project, we see a very different trend. We can now see that there are over 88,222 instances of HES organisations participating in a project, followed by 73,776 PRC participations and 57,192 from REC institutes. This shows that on average there are 29.4 projects per HES, while there are only an average 2.3 for each PRC. As the next figure shows, the reality is that the projects are dominated by a small number of institutions.

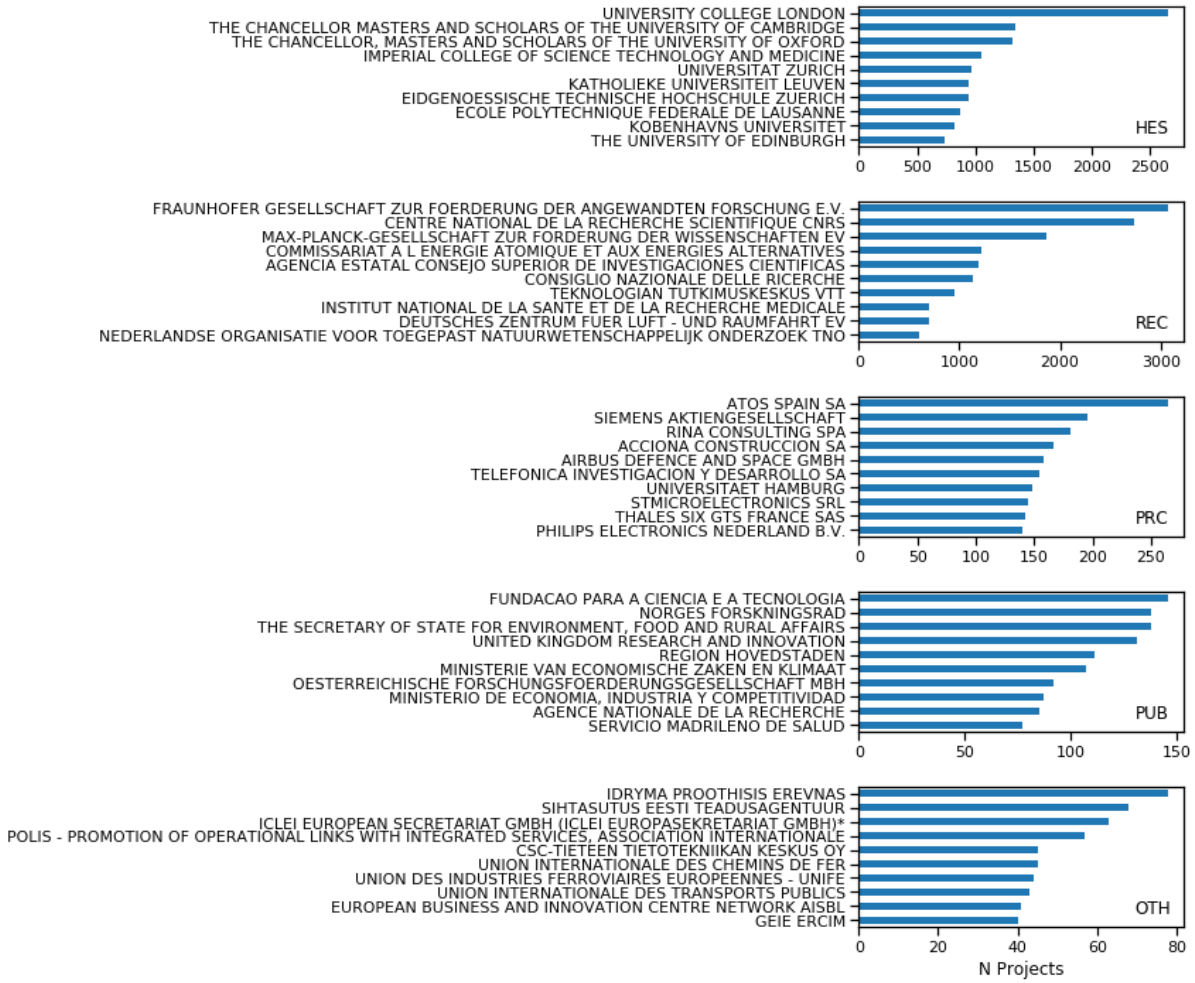


Figure 2.9: Top institutions by project participation count.

Here we see the top ten institutions by the number of FP7 and H2020 projects that they have participated in according to CORDIS. We can see that the participation rates in projects have a very skewed distribution, with a few actors accounting for the majority of projects. As an example, we can see that University College London is involved in over 2,500 of the 88,000 projects that involve a higher education establishment. The distribution is echoed across all institution types, though its exact numbers shape can be seen to vary.

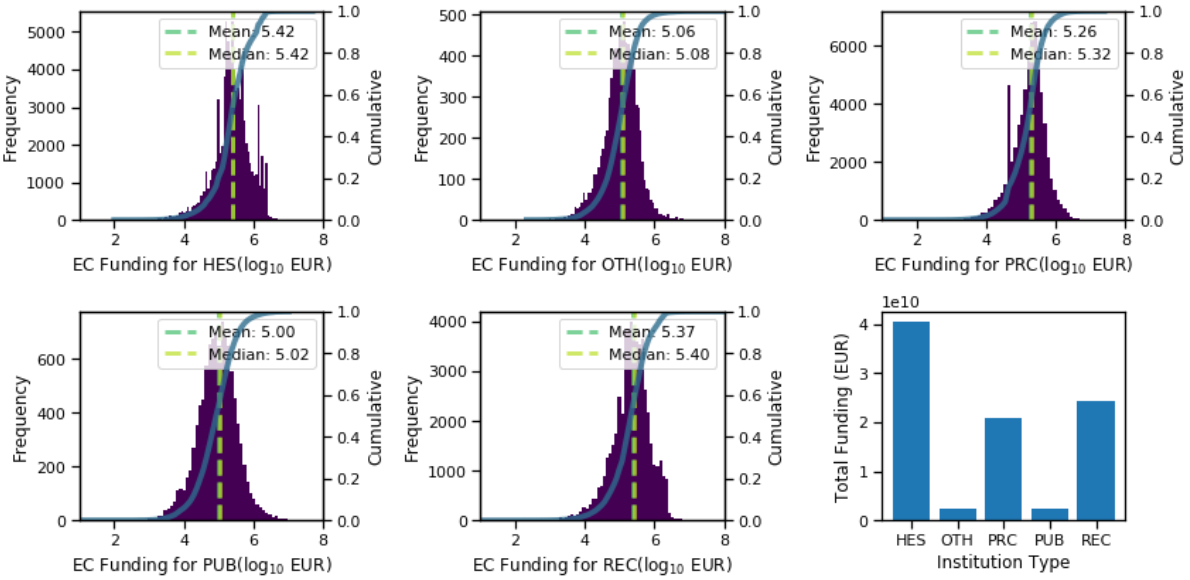


Figure 2.10: Funding by institution type.

CORDIS contains a breakdown of the funding received by each organisation within a project. While the most funding in total has gone to HES institutes, we can see that the distribution median of funds awarded to all types of organisations falls between the range of €104k and €263k.

Publications

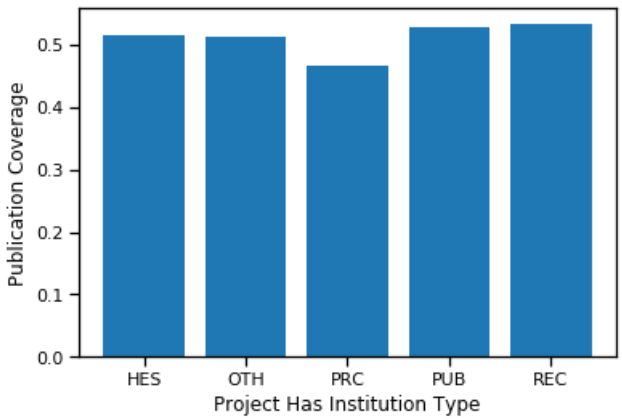


Figure 2.11: Coverage of CORDIS projects by OpenAIRE publications by organisation type.

The publications gathered from OpenAIRE do not have perfect coverage of the projects listed in CORDIS. It is unknown to what extent the lack of coverage is caused by projects not having published any outputs or simply that those outputs have not been ingested into the OpenAIRE database yet. The coverage is around 50% across all projects, with the breakdown for projects involving at least one of a particular institution type shown above. Although there is a small dip for projects involving a private organisation, the coverage is around 50% for projects involving all organisation types.

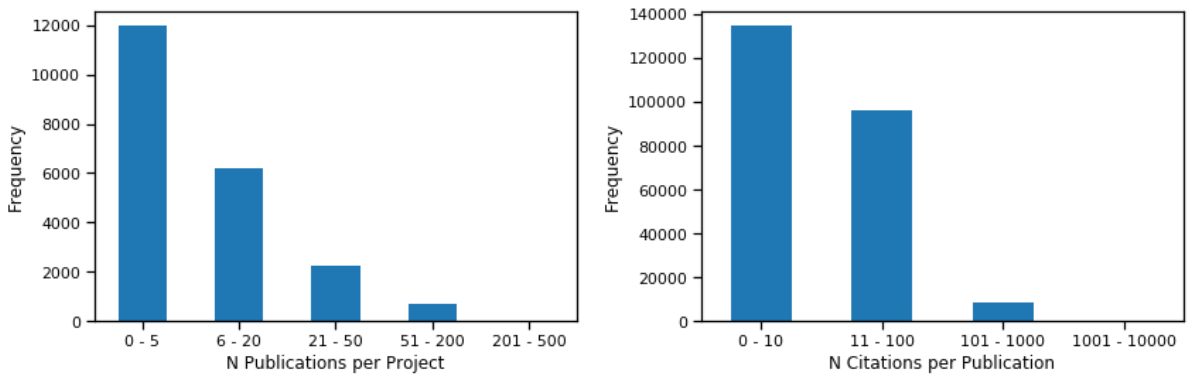


Fig 2.12. Number of publications and citations from FP7 and H2020 projects.

The numbers of publications per project show us that around 50% of projects publish 5 papers or less, while a very small fraction publish more than 200. From these publications, we can see that only a very small fraction achieve citation counts of greater than 100. These are the kinds of distributions we would expect, however it is interesting to know the effect of the number of contributors on the number of outputs from a project.

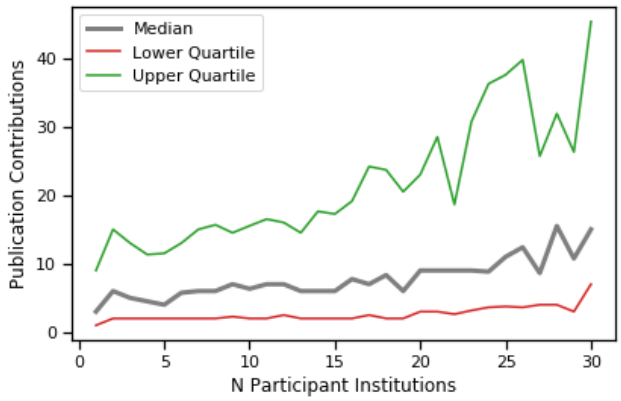


Figure 2.13 Average number of publications by number of project participants.

Grouping projects by their number of participating institutions, and plotting the median number of publications for projects in each group, we can see that there is indeed a slight upward trend. No values for projects with more than 30 participants are shown as the number of projects becomes low and the results are noisy.

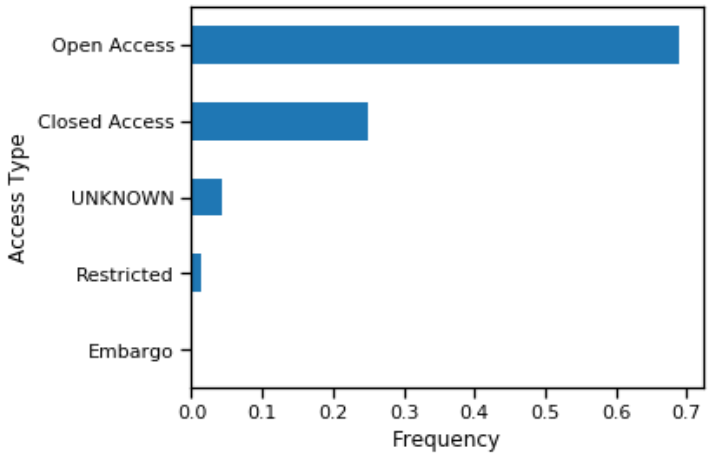


Fig 2.14. Access types of publications from FP7 and H2020 projects.

In the OpenAIRE database, each publication title has associated 'child' publication instances. These represent each appearance of a publication in a repository that is captured by OpenAIRE. Each of these has an associated access type that denotes whether the publication instance is open access, closed access, restricted, or under embargo. For each publication, we count whether any of the access types are present in its child instances, sum these counts, and then divide by the total number of child instances. This shows us that almost 68.9% of the publications have at least one open access form, while 24.9% have a closed access instance. A very small fraction are unknown, restricted or under embargo. Further calculations tell us that only 3.3% of publications have instances with access that is both open and closed.

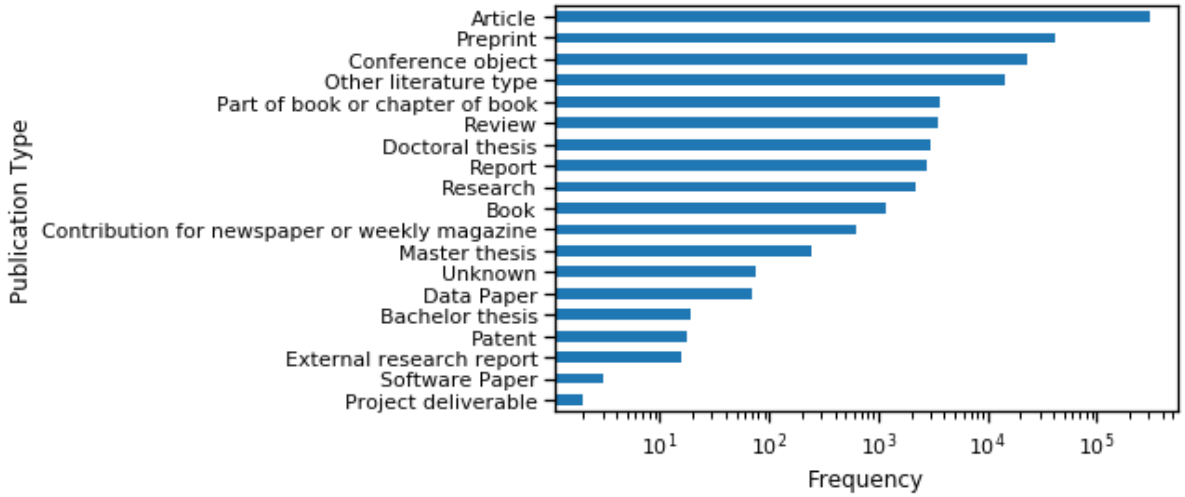


Fig 2.14:. Publication types from FP7 and H2020 projects.

In addition to telling us the access type for each publication instance, OpenAIRE also records what type of output each instance is. Counting these across projects shows that articles, preprints and conference papers are the most frequently occuring types, though a significant number are also listed as an unknown type. The two questions we ask from here are what is the distribution of access types are among those publications, and whether the distribution of their frequencies varies across subjects.

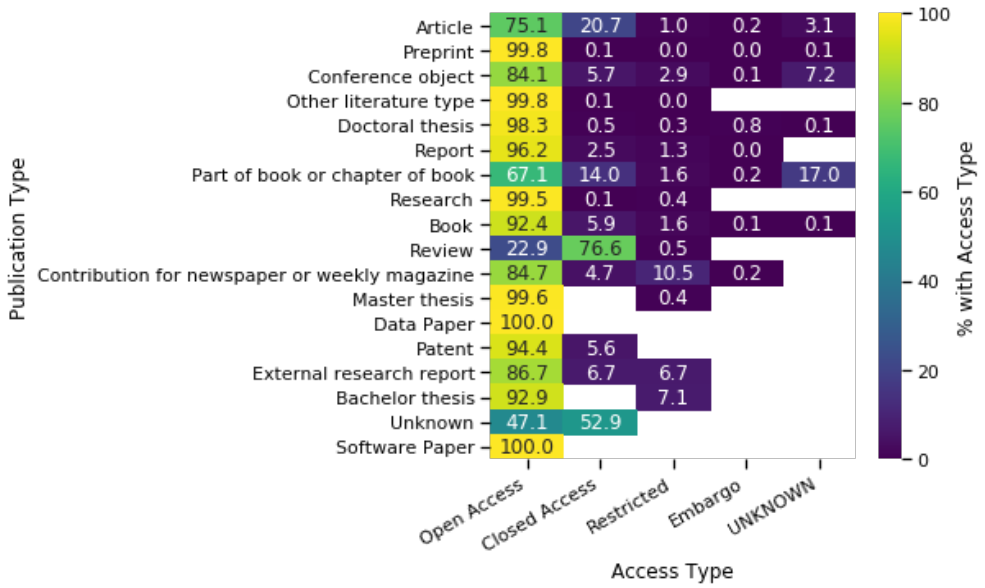


Figure 2.15: Normalised distributions of access type by publication type.

We can see that for 11 out of 18 publication types, over 90% of the outputs are open access. However, for articles, the most frequent type of publication, we can see that over 20% are closed access, which explains the high count of closed access publications seen in Fig. 2.14. It is only for publications of type review and unknown where the fraction of closed access instances exceeds that of open access, though book chapters, books, patents, external research reports and conference objects also have a small but notable fraction of closed access outputs. Contributions to newspapers or weekly magazines exhibit a 10.5% rate of restricted publication instances, which makes sense due to paywalls on these types of publications. 17% of book chapters and 7.2% of conference objects are classified as having an unknown access type, and the reason for this is not clear from the data.

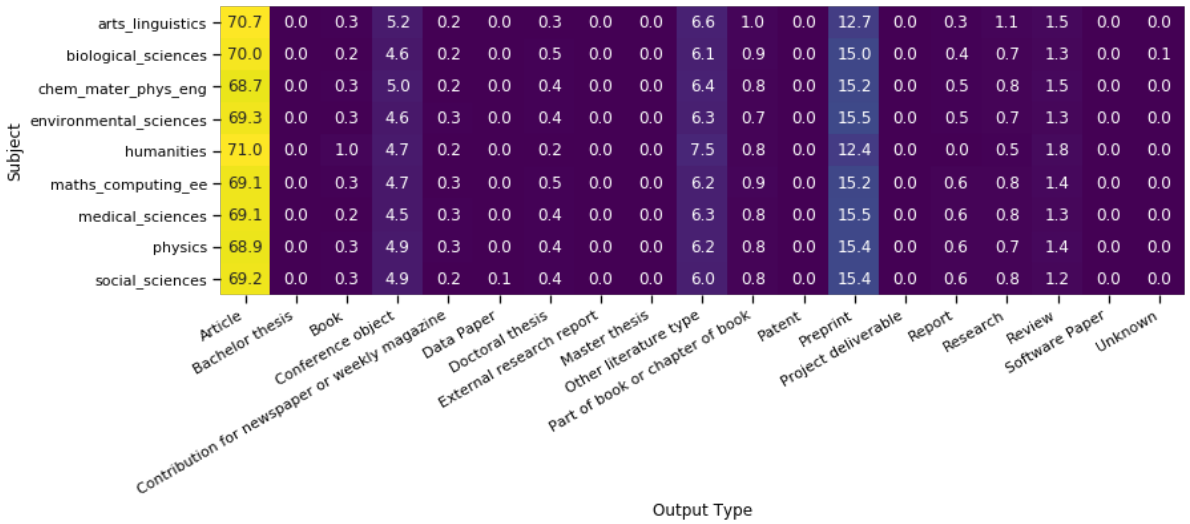


Figure 2.16: Percentage of publications by type and subject.

For papers with abstracts available, we are able to classify them into subject areas and then break down the percentage of publications by type within them. There are a few notable trends here that we might expect. First that we see that the arts and humanities are the only two subjects with preprint fractions below 15%, at 12.7% and 12.4% respectively. Arts and linguistics have a slightly higher fraction of

conference objects (perhaps due to the overlap between linguistics and computer science) and other literature types and ‘research, while the humanities have a slightly larger fraction of books and reviews. The sciences generally have slightly lower fractions of articles in favour of more preprints and are generally lower for books. While we can see these trends, the differences between disciplines are small and perhaps not as pronounced as we had previously assumed. This may reflect the nature of the data collection or that the subject classification is not granular enough to pick up the more pronounced practices within disciplines.

3.2 Findings

In this section, we investigate the creation of 3 possible advanced metrics that make use of the data explored in this pilot. With each of them, we strive to demonstrate a different dimension of the outputs from a project. The first is a normalised measure of citation count for a project. This metric intends to demonstrate a traditional bibliographic impact measurement that also takes into account subject area, time and resources allocated to a project. We then have a measure of novelty, that exploits the use of cooccurrence networks based on publication keywords provided by MAK. This shows the use of a new technique for metric creation that takes advantage of available web data. Finally, we show the linguistic similarity between a project’s original objectives and its reporting summaries or publication abstracts. This is used to demonstrate the possibility of turning qualitative unstructured data into a quantitative metric.

3.2.1 Normalised citations

Citations have long been used as a measure of impact of a piece of work on the scientific community. However simply counting them leads to significant biases: older publications will have had more time to accumulate citations; citation rates differ between disciplines; the absolute number of references made between publications changes over time. To create a metric of normalised citations for a project, we must take these factors into account. In addition, when considering the number of citations we can attribute to a project, we must consider that each publication may be the product of work from more than one funding award. Indeed, the OpenAIRE data contains many publications that are linked to two or more European Commission project codes.

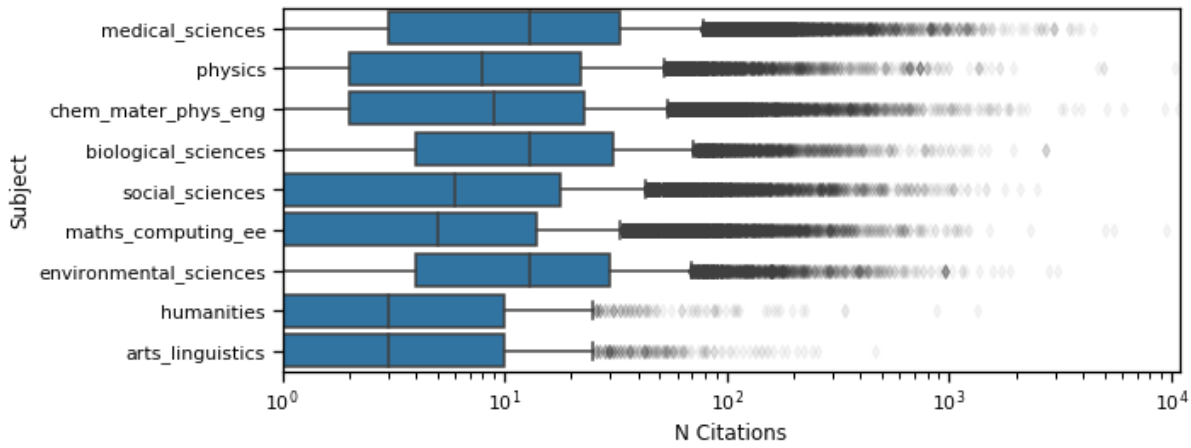


Figure 3.1 Distribution of citation counts for publications by field.

By examining the distributions of citation counts, we can see that the median value for most fields is on the order of 10^1. However, there are also publications with many thousands of citations in the database. We can also see that there is a significant difference between the citation rates in the different fields.

First, we normalised the number of citations for each paper by the total number of citations for all publications within each year, to account for overall citation trends. We then account for the age of each publication by dividing the year-normalised citations for each publication by its age, which we calculate as the difference in time between the date that the data was collected and the its publication date. We also divide by the number of citations by the author count on each paper. Finally we, attribute a portion of the citations for each publication to a project. For this work, and with the lack of any information that proportionately attributes projects to publications, we make the assumption that for a publication that has more than one project code listed, every project contributed equally to the work undertaken. In practice, this means that we divide the citations equally between the projects on multi-project publications. Finally, we take the $\log_{10}$ for each normalised citation count, in order to reduce the range of the spectrum. As we have seen, it is possible to have citation counts that vary by many orders of magnitude, but this is not very interpretable. By applying a log transform, the scale becomes compressed.

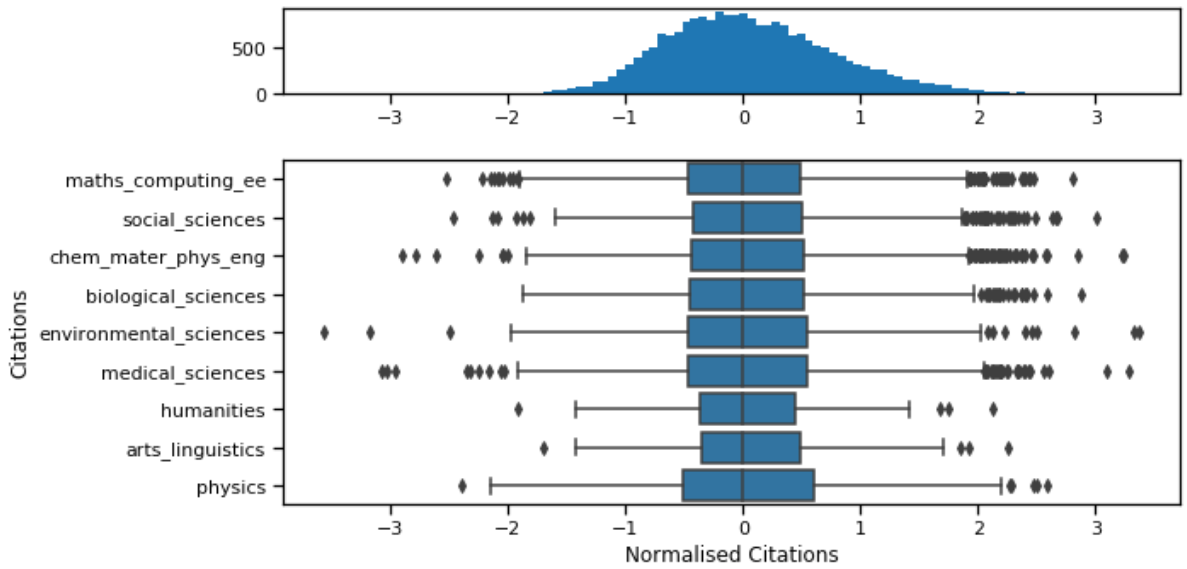


Figure 3.2: Normalised citation count by field.

We can see from the normalised citation count that we now have a distribution for each subject in which the median is centred on zero. We can also generally see that for the sciences, the width of normalised citation count distributions is fairly similar, and likewise for the humanities and arts and linguistics. We can see that the overall distribution of normalised citations is a very slightly skewed normal distribution.

3.2.2 Linguistic similarity

The CORDIS database contains objectives for each project, written as part of the original funding proposal, and progress reports which written as the project progresses or is completed. These unstructured free text data offer a qualitative source of information, which we can also take advantage of to form a quantitative indicator by using modern natural language processing methods.

Here we measure the difference in the use of language used during a project objective and project progress reports. To calculate the document similarity, we train a Doc2Vec model on all of theses texts, and then compare the document embeddings for the object of each project to the embedding for each report with the same project code with the cosine similarity. For each project, we then take the mean cosine similarity as our final value.

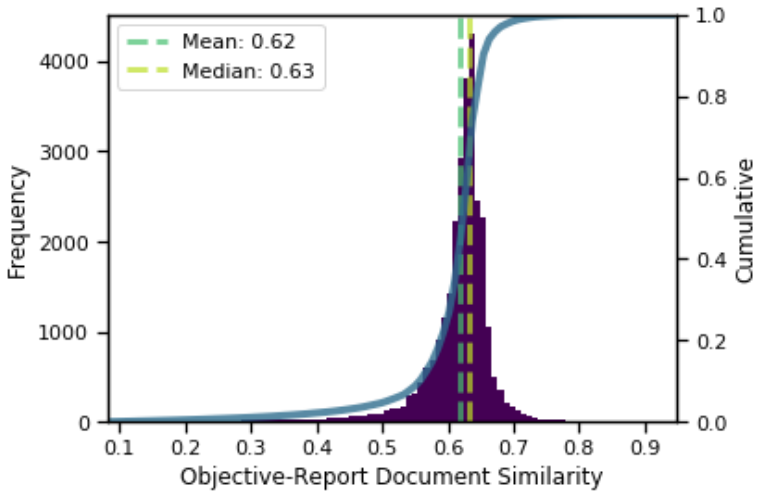


Figure 3.3: Document similarity between project objectives and report summaries.

As we can see, most documents are clustered around a similarity of the mean value 0.63, while the range spans values below 0.1 to values close to 1. From this graph alone however, the nature of the similarities between documents is not clear. To illustrate the differences being captured, the first 1,000 characters of the top 5 most similar and top 5 most dissimilar document pairs are presented in Appendix A. One issue that viewing the actual document pairs presents is that the original objectives, or snippets of them, are sometimes quoted in the reporting documents. This presents a challenge as the document similarity will be inflated for projects where this is the case. Further work is required on language filtering and normalisation to overcome this challenge.

3.2.3 K Factor

Each publication in MAK is tagged with a series of field names. Microsoft used network analysis of subjects on Wikipedia to identify around 200,000 fields, and has used natural language processing and deep learning to apply these fields to publications. The fields include high level disciplines such as ‘physics’, but also very granular topics such as ‘internet meme’. Here we take advantage of this labelling, and use it to create an indicator of novelty for each publication.

The underlying idea behind this indicator is that papers which exhibit a combination of fields that has not been seen before are more novel than those which have a combination of fields that has appeared in many previous publications. For example, a paper the has been labelled with the fields ‘particle physics’

and ‘boson’ is likely to be scored lower than a paper that is labelled with ‘particle physics’ and ‘chocolate’.

The method begins with creating a cooccurrence network of fields in all papers at time intervals which across the period covered by the set of publications being analysed. In this study, we use the interval of one year. The cooccurrence networks have nodes which represent the fields and edges that represent the cooccurrence of two fields in at least one publication. The edges are then weighted with the total count of cooccurrences between two fields in that year. For each year, we then take the cumulative sum of edges and weights for all previous years, so that the network at each year contains information about the relationships between fields for that year and all time previous.

With these cumulative cooccurrence networks, we are able to then calculate the added contribution that each publication has made to the overall network. We take a publication at time T, and generate the pairwise cooccurrences of the fields that it has been labelled with. We then compare this to the overall cooccurrence network from T-1, by finding the cumulative cooccurrence counts for each field pair, and then calculating the percentage increase that is gained by adding another single cooccurrence to each edge. This percentage gain is thought of as the knowledge enhancement that is gained by the addition of a link between two fields, or the K factor. Each publication may have several pairs of cooccurring fields, and so the K factor for a publication is taken to be the mean of the percentage gains exhibited for each pair of fields that it has been tagged with.

While this doesn’t capture the very specific novelties that may be exhibited within a publication, it is aimed at capturing novelty at the field level, and is higher for publications that bring together disparate areas of study. However, this indicator is not designed to create a polarising scale between novel and routine work. As it works at the field level, it is more indicative of whether a project is opening up new areas of research, or providing contributions to already established domains. The quality of these contributions can only be determined by using this measure alongside other indicators.

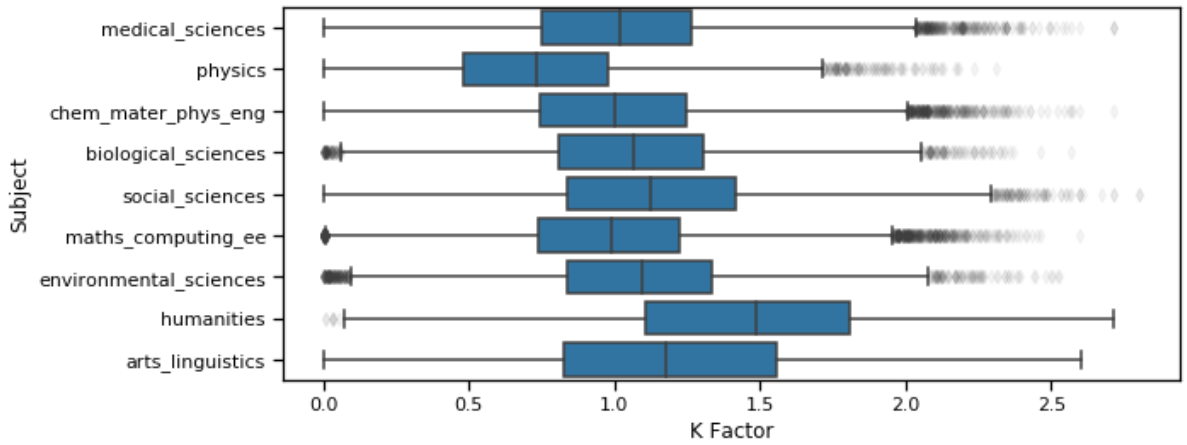


Figure 3.: K factor for publications in each subject.

The K factor was calculated for every publication. In addition to calculating the absolute value, it was normalised for each publication against the mean of all other publications released in the same year. It is also important to note that the K factor in this case does not represent the gain in disciplinary novelty compared to all research, but rather compared to other papers from EU projects funded and carried out within the same time period (2007 to 2019).

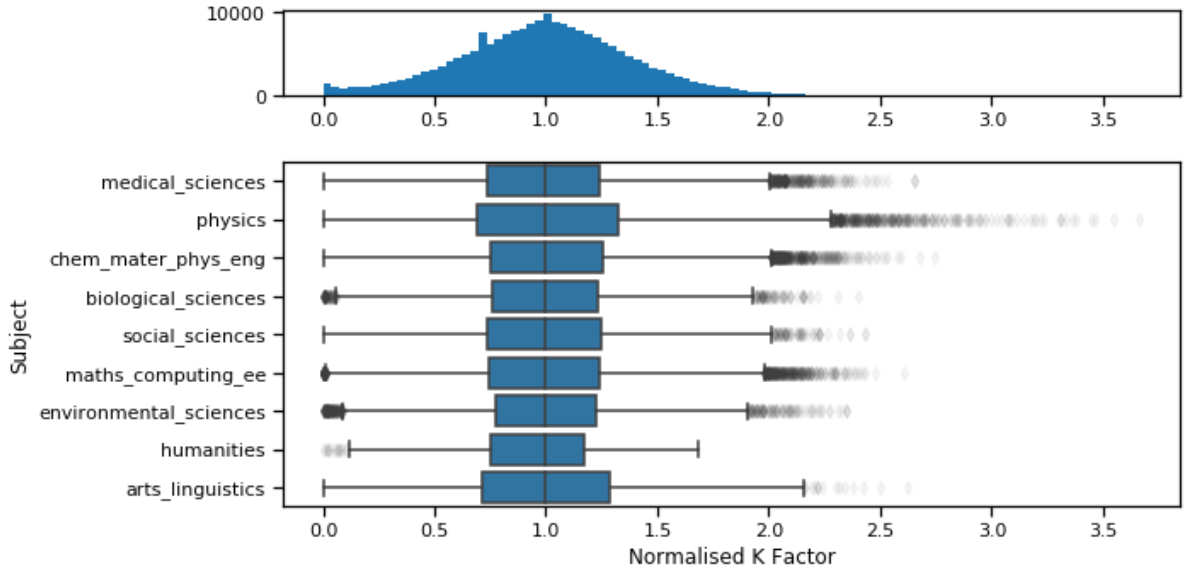


Figure 3.: K factor for publications normalised by subject.

To account for differences in K factor between disciplines, we normalise again by the median value for each subject classification, so that fair comparisons can be made between publications across all research. We can now see that each discipline has publications with a range of K factors from close to zero, up to 3.5. The distribution is close to normal on a linear scale, allowing us to make reasonable and interpretable comparisons between publications. We do see some spikes at around 1 and 0.75. This is perhaps due to publications that have been labelled with low numbers of fields, or noise from years with few publications. A list of titles with high K factors in each subject area can be found in Appendix B.

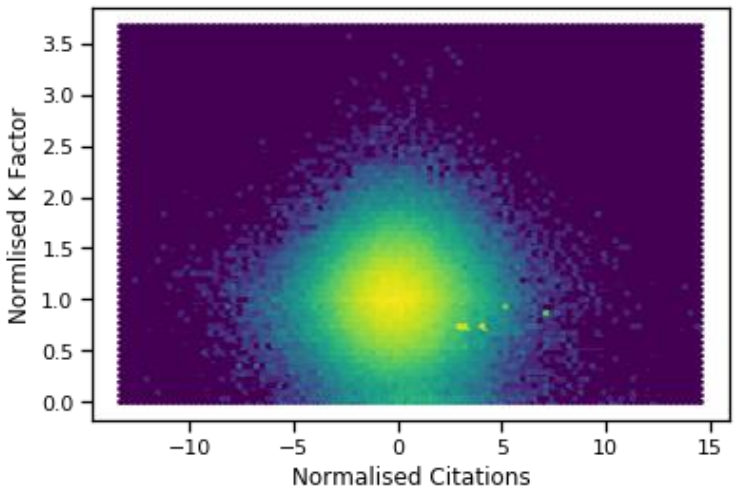


Figure 3.: Normalised K Factor vs normalised citation count.

What is important to note, is that the fully normalised K factor does not correlate with the normalised citation count. This shows that this indicator is measuring something different from the amount of attention that a paper has received in the form of references. This makes sense, as many papers may create new links between fields, but go relatively unnoticed by the community, the audience for such publications may not yet exist, or the quality of the work may not be as high as that of others which are cited more frequently.

4 Discussion and Conclusions

This pilot has presented preliminary analysis of data from CORDIS, OpenAIRE, Crossref and MAK, as well as examples of 3 possible indicators that might form a suite of advanced RITO indicators for research and innovation evaluation.

We have seen that the case for a set of indicators that can capture many dimensions of research and innovation inputs and outputs are justified, as it is clear that the EU funds a wide range of projects. There is a broad distribution of the size of projects, including funding amounts, project durations, numbers of participant organisations, and the number of countries involved. The portfolio of projects also spans the public and private sector, with the largest slice of funding being awarded to higher education establishments and research institutions, but a significant portion also going to private organisations.

Using a list of publications from EU projects hosted on OpenAIRE, and collecting publication metadata for them from Crossref and MAK, we were able to explore the nature of published outputs from the FP7 and H2020 portfolio. We found a range of publication types and under different levels of access, though the majority are open access journal articles or preprints. There are subtle differences between the publishing patterns for projects in different disciplines, particularly between the sciences and the arts and humanities. While this only represents project outputs in terms of published results, it already shows that there is a spectrum of dimensions of impact that could be measured.

We then focused on creating 3 indicators that exhibited some of these dimensions. This includes a normalised measure of citation count, a measure of linguistic similarity, and a measure of subject area novelty. While citations area already used as a measure of quality for research publications, they are often taken at face value. For example, the top publications from a set of projects might be identified as those that are within the top 1% of all publications by citation count. Here we introduce an indicator that normalises the citation count by a number of factors that will influence the absolute count; the publication's age; the number of authors; the subject area; overall citation trends. This provides a metric that seeks to communicate publication quality with respect to the context that it was created within. This metric can be seen as complementary to the absolute citation count. While it provides an improved method to compare publications across the entire portfolio, it is still useful to know the absolute numbers of citations as this can also tell us about the overall reach and impact of a particular publication.

We then presented a measure of document similarity between project objectives and progress reports in CORDIS. This is intended to demonstrate the possibility of assessing the difference in language between the start of a project and the waypoints that it passes. However, the usefulness of this indicator is currently fairly limited without further cleaning of the text and accounting for common occurrences such as the original objectives being directly quoted within the progress reports. As with other indicators, it should be normalised by subject area.

Finally, we create a new indicator, named the K factor, which use a new data source to calculate an indicator that represents research novelty. The K factor uses keywords from a publication, in this case field names from the MAK API, and measures to what extent they form a novel combination of research areas in comparison to previous research. A dynamic graph of the interconnectedness of knowledge within the set of research papers is created by calculating cumulative cooccurence graphs of research fields over a period of time. The combination of fields from each publication within every time period is then compared to the cumulative sum of knowledge cooccurrence in previous periods, in a bid to find

how many concepts have been newly linked by the publication. In essence this is a measure of how much a publication is opening up new interdisciplinary research areas, versus how much it is contributing to established research pathways.

4.1 Validation and ongoing stakeholder engagement

As highlighted in the introduction, indicators should be designed as a tool for research and innovation assessment that is complementary to peer review and expert judgement. To validate the methods of indicator creation, and to create new indicators in the same vein, it would be advantageous to seek the input of domain experts in two ways. It would be useful to understand to what degree stakeholders' judgements of research agree with the indicators that are being created. This could be done through the form of surveys, asking academics or policy makers to judge different dimensions of research projects or outputs. Additionally, it would be useful to have the feedback from stakeholders on the usefulness of the indicators presented here, and to understand what else they might consider useful after having seen these.

4.2 Limitations

There are of course limitations to the work displayed here. The most prominent is the reliance on publication data to assess project outputs. While a range of publication types are captured, they cannot possibly cover all of the possible outputs and impacts that might be achieved by a piece of research. Publications are also more focused on academic work, and further materials would be needed to assess innovation happening outside of public research establishments. Even within more academic research, the work here misses conference presentations, public engagement, and other activities through which an academic may disseminate their findings. Additionally, the coverage of publications to projects was only around 50%. It is not clear what proportion of this is due to projects having not yet published (and some may never do so) or data coverage.

The methods presented here do not address the issue of measuring long term impact. This remains challenging to measure as it is hard to track the influence of a single piece of research or innovation over time. This is an issue due to limitations in linking data across long timer periods, and also the fact that there are many confounding factors.

Although we present a method of citation measurement that normalises across subject areas, it is not clear whether differences in citation patterns between disciplines are a 'natural' occurrence of the result of incentives created by previous policy decisions. This highlights that in addition to creating new indicators, it is important to analyse feedback loops that they might be responsible for creating.

For our document similarity metric, we use Doc2vec, which creates a quantitative measure based on qualitative information. However, this should only be used illustratively to show an evolution in language, rather than as a marker of research quality. It is not a trivial task to use natural language processing on completely free text such as research objectives and progress reports to determine whether the original aims of the research proposal have been met. This is a clear example where the balance of human-computer judgement is critical.

Our calculation of the K factor relies on machine labelling of publications with the research fields to which they apply. There are multiple issues of trust that arise from this. Researchers and policymakers may not trust the accuracy of machine powered labelling, or they may not trust the completeness of the

taxonomy used to do the task. While tags applied by machine are used for this study, we can imagine that in a real world scenario, the tags might be applied by the authors themselves, or reviewers, or by a machine and then validated by a human. This might ensure greater accuracy and trust in the methodology.

4.3 Considerations for scaling up

To truly create a suite of indicators that is able to measure the diversity of research and innovation project inputs and outputs, a wider array of data should be collected. One source could be text data scraped from the websites of companies involved in EU projects to understand the products being created by private organisations. Additionally, we might use the Crossref 'events' API endpoint to find out about the wider reach of academic outputs, such as mentions in the media, on blogs or on social media. This would capture dimensions that cover public engagement and consumption of research beyond academia. Yet another dimension that should be captured is geographic diversity of participants in research projects. Although the CORDIS data contain country level information, it does not provide exact locations for the project participants. This means that it is not possible to get a true sense of the distances between collaborators on a project, to understand whether this has an impact on outputs. While we were assessing outputs at the project level, the projects are carried out by individual researchers, and their level of overall or subject-specific experience may also play a role in project outputs. However, as mentioned earlier in this report, this is not an easy task, and would require further work to form a robust dataset.

Other analysis methods may also be incorporated to understand the research funding landscape and the impacts of projects more accurately. While this pilot has explored the possibility of creating new metrics, it has not yet investigated the relationship between them and the inputs to a project. If the purpose of new indicators is to aid future decision making around programmes and funds, then this will be a necessary step. The techniques to do this might range from simple univariate correlation analyses, to multivariate supervised machine learning methods. Additionally, we may want to explore clustering projects to get a more comprehensive understanding of the dimensions at play. Finally we would want to compare any new metrics and the results of models built upon them to the traditional metrics that are currently used such as citation counts, h-index or journal impact factor.

Communicating research and innovation indicator results

In a scenario where a suite of indicators that meet the RITO criteria has been created it is also important to consider how they might be visualised and communicated to stakeholders. Visualising highly multi-dimensional information in an interpretable manner is not a trivial task. Additionally, it is not the role of those creating the metrics to define what a succesful or 'good' project looks like, therefore any communication strategy should provide a neutral portal through which a decision maker can access the information required to answer their own needs. One potential option may be to highlight groups of similar research within an interactive data visualisation that allows a user to both focus in on a particular set of projects, or compare projects more broadly. We take inspiration from a recently published data driven visualisation of countries, which compresses many country level statistics into a two-dimensional layout, and uses colour and a control bar to enable the user to easily switch between different views and groupings of nations. It is conceivable that a similar approach might be employed to view research projects and obtain a high level overview of the features that characterise them.

Figure 4.1: An alternative data-driven country map.

4.3.1 Complementarities with other pilots

This pilot is complementary to the inclusive innovation pilot. The indicators of inclusivity should also become a part of the research and innovation indicators, either as input features, to highlight the impact of diversity on research outcomes, or as project outputs, to show the effects of different project types on enhancing inclusivity across research and innovation.

4.3.2 Tools and data sources

This pilot uses Python 3.6. All the analysis was performed in memory on an 8GB 2.4GhZ laptop and carried out and documented in JuPyteR notebooks. An online repository (https://github.com/nestauk/funding_analytics_eu) contains draft versions of the notebook still to be refactored and documented.

5 References

Belikov, Aleksey V., and Vitaly V. Belikov. 2015. “A Citation-Based, Author- and Age-Normalized, Logarithmic Index for Evaluation of Individual Researchers Independently of Publication Counts.” *F1000Research* 4 (September): 884. https://doi.org/10.12688/f1000research.7070.1.

Cagan, Ross. 2013. “The San Francisco Declaration on Research Assessment.” *Disease Models & Mechanisms* 6 (4): 869–70. https://doi.org/10.1242/dmm.012955.

Dunleavy, Patrick. 2011. “The Research Excellence Framework Is Lumbering and Expensive. For a Fraction of the Cost, a Digital Census of Academic Research Would Create Unrivalled and Genuine Information about UK Universities’ Research Performance.” *Impact of Social Sciences* (blog). June 10, 2011. https://blogs.lse.ac.uk/impactofsocialsciences/2011/06/10/ref-alternative-harzing-google-scholar/.

Gibney, Elizabeth. 2016. “Major Review Calls Time on ‘gaming’ in UK Research Assessment.” *Nature News*, July. https://doi.org/10.1038/nature.2016.20343.

Guthrie, Susan, Joachim Krapels, Catherine A. Lichten, and Steven Wooding. 2016. “100 Metrics to Assess and Communicate the Value of Biomedical Research.” Product Page. 2016. https://www.rand.org/pubs/research_reports/RR1606.html.

Guthrie, Susan, Watu Wamae, Stephanie Diepeveen, Steven Wooding, and Jonathan Grant. n.d. “Measuring Research: A Guide to Research Evaluation Frameworks and Tools,” 7.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. “Bibliometrics: The Leiden Manifesto for Research Metrics.” *Nature News* 520 (7548): 429. https://doi.org/10.1038/520429a.

Lane, Julia. 2010. “Let’s Make Science Metrics More Scientific.” *Nature* 464 (March): 488–89. https://doi.org/10.1038/464488a.

Martin, Ben R. 2011. “The Research Excellence Framework and the ‘Impact Agenda’: Are We Creating a Frankenstein Monster?” *Research Evaluation* 20 (3): 247–54. https://doi.org/10.3152/095820211X13118583635693.

Publications Office of the European Union. 2015. “Horizon 2020 Indicators : Assessing the Results and Impact of Horizon.” Website. November 17, 2015. https://publications.europa.eu/en/publication-detail/-/publication/68686e76-8f53-11e5-983e-01aa75ed71a1/language-en.

Sardesai, Ann Veena. 2014. “An Investigation of the Impacts of Excellence in Research for Australia: A Case Study on Accounting for Research.” PhD, Queensland, Australia: Queensland University of Technology.

Smith, Simon, Vicky Ward, and Allan House. 2011. “‘Impact’ in the Proposals for the UK’s Research Excellence Framework: Shifting the Boundaries of Academic Autonomy.” *Research Policy* 40 (10): 1369–79. https://doi.org/10.1016/j.respol.2011.05.026.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, et al. 2015. “The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.” Unpublished. http://rgdoi.net/10.13140/RG.2.1.4929.1363.

6 Appendix

Appendix A:
CORDIS Objectives and reports with the highest document similarity.

##########
Similarity: 0.95
=== Objective ===
Over the last decade, Human-Computer Interaction (HCI) has grown and matured as a field. Gone are the days when only a mouse and keyboard could be used to interact with a computer. The most ambitious of such interfaces are Brain-Computer Interaction (BCI) systems. BCIâs goal is to allow a person to interact with an artificial system using brain activity. A common approach towards BCI is to analyze, categorize and interpret Electroencephalography (EEG) signals in such a way that they alter the state of a computer. ACoBSECâs objective is to study the development of computer systems for the automatic analysis and classification of mental states of vigilance; i.e., a personâs state of alertness. Such a task is relevant to diverse domains, where a person is required to be in a particular state. This problem is not a trivial one. In fact, EEG signals are known to be noisy, irregular and tend to vary from person to person, making the development of general techniques a very difficult sc ...
=== Report ===
Over the last decade, Human-Computer Interaction (HCI) has grown and matured as a field. Gone are the days when only a mouse and keyboard could be used to interact with a computer. The most ambitious of such interfaces are Brain-Computer Interaction (BCI) systems. BCI's goal is to allow a person to interact with an artificial system using brain activity. A common approach towards BCI is to analyze, categorize and interpret Electroencephalography (EEG) signals in such a way that they alter the state of a computer. ACoBSEC's objective is to study the development of computer systems for the automatic analysis and classification of mental states of vigilance; i.e., a person's state of alertness.
Such a task is relevant to diverse domains, where a person is required to be in a particular state. This problem is not a trivial one. In fact, EEG signals are known to be noisy, irregular and tend to vary from person to person, making the development of general techniques a very difficult scient ...

##########
Similarity: 0.93
=== Objective ===
The Programme for Post-Doctoral Talent Attraction to Campus do Mar, FELLOWSEA, is a programme developed by the University of Vigo (UVIGO) which intends on building particularly favourable environment to attract the best experienced researchers in marine and maritime research. Within the proposed 48 months co-funding period (36 months of fellowship), FELLOWSEA will offer to 10 incoming experienced researchers, non-Spanish residents, the opportunity to develop her/his research project within the framework of a vast network of institutions participating under Campus do Mar, an International excellence initiative in marine and maritime themes for the service of Society.
Fellows will have full trans-national mobility experience accessing to research facilities of the partners FELLOWSEA, and been hosted by one of the several research groups existing in the hosting institutions: the University of Vigo (UVIGO), the University of Santiago de Compostela (USC), the University of A CoruÃ±a (UDC), ...
=== Report ===

The Programme for Post-Doctoral Talent Attraction to Campus do Mar, FELLOWSEA, is a programme developed by the University of Vigo (UVIGO) which aims to build particularly favourable environment to attract the best experienced researchers in marine and maritime research. Within the proposed 48 months co-funding period (36 months of fellowship), FELLOWSEA offers 9 incoming experienced researchers, non-Spanish residents, the opportunity to develop her/his research project within the framework of a vast network of institutions participating under Campus do Mar, an International excellence initiative in marine and maritime themes for the service of Society.

As of 2015, 9 researchers were recruited. Fellows have full trans-national mobility experience accessing to research facilities of the partners FELLOWSEA, and are hosted each by one of the several research groups and research centers existing in the University of Vigo (UVIGO). Researchers were selected through a transparent and inter ...

##########
Similarity: 0.87
=== Objective ===
The objective of this project is to uncover and explain the escalation and non-escalation of repression and intra-state armed conflict by analyzing how characteristics of the government and its formal and informal security apparatus shape the dynamics of such violence, paying particular attention to the role of monitoring and accountability. RATE analyzes when and under what conditions what types of human rights violations lead to the escalation or deterrence of further repression and armed conflict. Although there has been substantial increase in research on civil war, we know surprisingly little about the dynamics that escalate armed conflict within country-borders and those that prevent an escalation and what role human rights violations and informal armed actors play in those dynamics. While civil wars are a relatively rare occurrence, repression and human rights violations are not. What can this tell us about the link between human rights violations and repression? What leads to t ...
=== Report ===
This project analyzes when and how political violence committed by the state escalates. Our findings from global cross-temporal studies show that when the government limits the freedom of the media, it is likely to subsequently escalate its physical repression of the wider population. These results highlight how the respect for civil liberties and physical integrity rights are closely intertwined and how the violation of press freedom is often a precursor for more government violence.

Using a novel dataset that utilizes information collected by three globally active non-governmental organizations, we show that the killing of journalists acts as an early warning signal for deteriorating human rights conditions in the following two years. It emphasizes the importance of closely monitoring the safety of journalists, not only for the security of the journalists but also as an indicator of increasing government violence. Using our novel data we provide new insights into where most journ ...

##########
Similarity: 0.87
=== Objective ===
There is worldwide consensus that the e\e- International Linear Collider (ILC) is the next major project in High Energy Physics following the imminent commissioning of the LHC; it is a high priority in the European Strategy for Particle Physics agreed by CERN Council. The ILC will

constitute the precision tool for the Terascale, the scale of electroweak symmetry breaking. The ILC complements the potential of the LHC, which will initially chart this unknown territory. The ILC-HiGrade project brings together the key players in Europe to engage towards the realisation of the ILC. They constitute a large fraction of the European element of the Global Design Effort (GDE) that has recently led to the publication of the Reference Design Report (RDR). The report forms the basis for the Engineering Design Phase of the ILC, which the GDE will complete by mid-2010 when the proposal for the ILC will be presented to the global stakeholders, i.e. governments and funding agencies to seek approval. Th ...
=== Report ===
Project context and objectives:

A linear e+e- collider continues to be the next major project in high energy physics (HEP) following the successful start of operations of the large hadron collider (LHC) and the first presentations of physics results. A linear collider has been prominently positioned in the European strategy for particle physics agreed by the European Organisation for Nuclear Research (CERN) council, which serves as the basis for the European strategy forum for research infrastructures (ESFRI) recommendations for HEP. The initial physics results emerging from the LHC give confidence that the field will receive more guidance from the LHC by the end of 2012 - when large statistics samples become available - on the detailed design decisions for such a linear collider, in particular the energy reach of such a facility. This information is timely for the update of the European strategy, which will be released at the end of 2012.

In the energy range from 500 to 1 000 ...

###########
Similarity: 0.87
=== Objective ===
The QB50 Project will demonstrate the possibility of launching a network of 50 CubeSats built by CubeSat teams from all over the world to perform first-class science and in-orbit demonstration in the largely unexplored middle and lower thermosphere. Space agencies are currently not pursuing a multi-spacecraft network for in-situ measurements in the middle and lower thermosphere because the cost of a network of 50 satellites built to industrial standards would be very high and not justifiable in view of the limited orbital lifetime. No atmospheric network mission for in-situ measurements has been carried out in the past or is planned for the future. A network of satellites for in-situ measurements in the middle and lower thermosphere can only be realised by using very low-cost satellites, and CubeSats are the only realistic option. The Project will demonstrate the sustained availability of low-cost launch opportunities, for launching small payloads into low-Earth orbit; these could be m ...
=== Report ===
Project Context and Objectives:
The QB50 Project will demonstrate the possibility of launching a network of 50 CubeSats built by CubeSat teams all over the world to perform first-class science and in-orbit demonstration in the largely unexplored lower thermosphere. Space agencies are not pursuing a multi-spacecraft network for in-situ measurements in the lower thermosphere because the cost of a network of 50 satellites built to industrial standards would be very high and not justifiable in view of the limited orbital lifetime. No atmospheric network mission for in-situ measurements has been carried out in the past or is planned for the future. A network of satellites for in-situ measurements in the lower thermosphere can

only be realised by using very low-cost satellites, and CubeSats are the only realistic option. The Project will demonstrate the sustained availability of a low-cost launch opportunities, for launching small payloads into low-Earth orbit; these could be micros ...

CORDIS Objectives and reports with the lowest document similarity.

##########
Similarity: 0.08
=== Objective ===
Scopio Labs aims to expand the use of Whole Slide Imaging (WSI) for its application in digital pathology and is developing a digital microscopy platform capable of scanning high resolution images with low hardware costs.
Despite digital microscopy exists for decades, its use has been limited by the high prices of the technology. Our solution provides a low cost-high quality technology which will allow the implementation of WSI facilities in small hospitals and diagnostic centres. By using cutting edge principles of computational photography, we are able to enhance resolution and magnification (up to 100X) of images taken by affordable optics and mechanics through a disruptive computation method. Biopsies are frequently used to determine different diseases as psoriasis, diabetes, hepatic diseases, infections or cancer. Every year, more than 3 million new cancers are diagnosed and cancer is causing around 1.7 million deaths worldwide. Its early diagnosis has been pointed out as a factor ...
=== Report ===
Speeding up, decentralizing and improving diagnostics is a major priority for national health systems in Europe as we face an aging population and emergent infections in our changing society. Microscope slide-based diagnosis (the standard for malaria, leukemia, pneumonia and many others) is currently analog, time-consuming and prone to human error. There is a recurring need (25% of the cases) of a second opinion, in which cases the slides must be shipped to dedicated centers. Digitization of slides addresses these problems by eliminating the need for physical shipment of glass slides, decentralizing, and speeding up the process from weeks to minutes. Digitization by Whole Slide Imaging technology is being promoted nationwide in many countries, but there exist no low-cost, high-magnification digital microscopes. This means digital pathology is not accessible to low-budget labs and hospitals, making diagnostics and treatment decisions take longer, disadvantaging, therefore, the patients. ...

##########
Similarity: 0.08
=== Objective ===
The overall objective of this 6-month Phase 1 project is to evolve the prototype of Biopsy X, the new reliable endoscopic biopsy instrument, taking multiple cancer samples of varying depth in one single session, to transform our initial business plan into a robust strategy for international commercialization and to prepare for a Phase 2 demonstration.

Our product is the missing link that will bring effective, fast and patient-friendly cancer diagnosis, setting up the new standard for this method. The EndoDrillÂ® Model X is a unique combination of knowledge in mechanical engineering, physics and physiology. Research has been partially funded by Swedish R&D grants from Vinnova as well as from foundations (Sten K Jonsson Foundation, IKEA Foundation) and Lund University Innovation. Since 2017 the company is listed at AktieTorget stock

market.
Our ambition is to accelerate proper diagnosis progress so that 3 in 4 people will survive cancer within the next 20 years by making sure patients ...
=== Report ===
With more than 3.7 million new cases and 1.9 million deaths each year, cancer represents the second most important cause of death and morbidity in Europe. Early and accurate diagnosis is a major societal challenge. Endoscopy with biopsy sampling are gold standard for diagnosis of some of the most common cancers. However, none of market available solutions has reached the required performance enabling substantial decrease in examination time and increased sampling accuracy. This unmet clinical need is a business opportunity that BiBBInstruments AB has identified and thus developed EndoDrill® Model X, a new innovative single-use medical instrument, enabling taking multiple samples in one session, without removing the instrument between every biopsy. Additionally, EndoDrill® Model X offers the possibility to sample both superficial and deeply situated biopsies, a unique feature in itself. BiBB`s innovation is beneficial for several reasons as it: saves healthcare resources (time and mo ...

##########
Similarity: 0.08
=== Objective ===
Modern farm tractors provide most of the muscle power needed for today's high output agricultural enterprises. However, with current practices, they are also dangerous, with a number of fatal occupational accidents estimated at three times higher than the average. Power Take-Off and contact with machinery attachments (known as hitching) are one of the main causes. A routine task in farming is to connect the tractor with a trailer or tool. Current approach is to do it manually, requiring the operator get near the connection, hence, the injuries.
We have developed the first fully automated hitching system, SIWI. With our system the operator can connect/disconnect the tractor from the seat, not having to get anywhere near the connection. Moreover, productivity is dramatically increased since a trailer change is done in just under 30 seconds! Instead of the current 15-20 minutes. Reduced cost, operation time and increased safety and comfort are SIWI USPs. Moreover, our solution is a great ...
=== Report ===
We validated the technical, commercial and financial feasibility of our business innovation project and established the plan for commercialising our breakthrough technology around the globe. We carefully analysed our target market and identified key market drivers that will fuel the market adoption of our products. To speed-up post-project sales, we fine-tuned our commercialisation strategy and contacted several distributors in our target markets. Technically, we refined our plan to improve the hydraulics system of our products and on fine-tuning the design of our new system that will be mounted on tractor’s transport block. Our market research focussed on analysing the global agricultural machinery industry, the needs of our target customers and elaborated on the value-added by our technology. We analysed key barriers and risks and established their mitigation actions. We studied our competition landscape in the hitch segment and our freedom to operate analysis validated that our tech ...

##########
Similarity: 0.08
=== Objective ===

Delivering on the 5G promise of increased data rates, and ubiquitous coverages, poses stringent requirements on traditional vertically integrated operators. In particular, telecom operators are expected to massively roll out Small Cells, which requires finding appropriate urban spaces with both backhaul and energy availability. Network sharing becomes essential to unlock those commercial massive deployments. The open access model, or neutral host, will come to play a key role on the deployment of 5G networks, especially in urban scenarios where very dense Small Cell deploymens are required.
In parallel recent trends are paving the way towards the development of new, heterogeneous and distributed cloud paradigms that significantly differ from todayâs established cloud model: with edge computing, cloud architectures are pushed all the way to the edge of the network, close to the devices that produce and act on data. We posit that there are two sets of players perfectly poised to take ...
=== Report ===
Delivering on the 5G promise of increased data rates and ubiquitous coverage poses stringent requirements on traditional, vertically integrated operators. In particular, telecom operators are expected to massively roll out Small Cells, which requires finding appropriate urban spaces with both backhaul and energy availability. Network sharing becomes essential to unlock this commercial massive deployment, and to ensure that such deployment is done in a cost effective way.

To tackle this, 5GCity introduces the concept of a neutral host, where municipalities deploy such infrastructure only once, and make it available to third-parties using cloud-like paradigms, where such parties can run smart city applications near the edge of the network, wherever and whenever it is needed. The main objective of the project is then to design and implement such a neutral host architecture, implement and deploy the platform needed not only to slice, control and manage underlying infrastructure but also ...

##########
Similarity: 0.09
=== Objective ===
Motor learning is a fundamental process enabling an organism to improve movement efficiency during a motor task. It is supported by motor cortex, which organizes movements into complex sequences. Yet, how this structure is informed of planned and generated actions is poorly understood. Anatomically, motor cortex is highly interconnected with the motor thalamus (Mthal). Interestingly, this structure is involved during the acquisition of different motor tasks, but also shows homologous roles in motor function to brainstem areas. In line with these observations, we hypothesize that pathways between the brainstem to Mthal represent an interesting and unexplored way for motor information to reach cortical areas during motor learning.
This project aims at exploring the anatomical organization and functional importance of brainstem-Mthal pathways in motor learning and transmission to the cortex. It will first explore the bidirectional synaptic organization of these pathways according to neuro ...
=== Report ===
The ability to learn and to adapt a motor program is of fundamental importance to all species, since it enables performance of complex movements, but also to adapt these learned movements to perturbation in the environment. However, the circuits involved in these processes are poorly known. An important brain structure in this context is the cerebellum, and in particular the deep cerebellar nuclei (DCN), which are the sole output of the cerebellum. Interestingly, the DCNs project strongly to the brainstem and to the motor thalamus, two major structures in motor function.
Indeed, premotor brainstem areas with projections to the spinal cord are a major locus needed for the

control of movement. A recent study has shown that different brainstem nuclei connect to motor neurons innervating forelimb and/or hindlimb muscles in very specific patterns. For example, the manipulation of glutamatergic neurons in the medullary reticular formation ventral part (MdV) in mice led to specific deficits ...

Appendix B: Titles and field names of top 5 articles by K factor for each subject classification.

Subject: arts_linguistics
==========
Title: cultural diffusion in humans and other animals
K Factor: 2.62
Fields: sociology, neuroscience, evolutionary biology

Title: necessary not sufficient the circulation of knowledge about stained glass in the northern netherlands 1650 1821
K Factor: 2.50
Fields: stained glass, small number, philosophy, literature, history and philosophy of science, craft, apothecaries system

Title: metrical systems of celtic traditions
K Factor: 2.43
Fields: welsh, terminology, poetry, old irish, irish, history, genetic relationship, classics, celtic toponymy, celtic languages

Title: a personal tour of cultural heritage for deaf museum visitors
K Factor: 2.38
Fields: visual arts, sociology, sign language, mobile device, interpreter, hearing disability, dissemination, cultural heritage, bespoke

Title: of shipwrecks and weddings borders and mobilities in europe
K Factor: 2.34
Fields: politics, movie theater, mobilities, literature, humanitarian crisis, drama, denunciation, cynicism, bride, art

Subject: biological_sciences
==========
Title: combining plant volatiles and pheromones to catch two insect pests in the same trap examples from two berry crops
K Factor: 2.41
Fields: tarnished plant bug, sex pheromone, semiochemical, raspberry beetle, miridae, lygus rugulipennis, byturidae, biology, anthonomus rubi, agronomy

Title: evaluation of the effects of space allowance on measures of animal welfare in laboratory mice
K Factor: 2.31
Fields: stocking, risk factor, perseveration, open field, laboratory mouse, economics, demographic economics, animal welfare, aggression

Title: microbial production of next generation stevia sweeteners
K Factor: 2.22
Fields: taste, sweetness, stevioside, steviol, rebaudioside m, rebaudioside d, rebaudioside a, glucosyltransferase, biology, biochemistry

Title: occurrence and potential causative factors of immune mediated hemolytic anemia in cattle and river buffaloes
K Factor: 2.19
Fields: medicine, immunology, immune mediated hemolytic anemia, coombs test

Title: pest categorisation of anthonomus bisignifer
K Factor: 2.18
Fields: weevil, rubus, rosaceae, phytosanitary certification, ornamental plant, medicine, horticulture, fragaria, cultivar, biotechnology, anthonomus

Subject: chem_mater_phys_eng
==========

Title: high value plant products from discovery to final product
K Factor: 2.75
Fields: process engineering, final product, business

Title: estrategias actuales de control de xylella fastidiosa
K Factor: 2.68
Fields: xylella fastidiosa, virology, history

Title: through container extremely low concentration detection of multiple chemical markers of counterfeit alcohol using a handheld sors device
K Factor: 2.59
Fields: tolerable level, spatially offset raman spectroscopy, scotch whisky, pulp and paper industry, environmental science, counterfeit

Title: durability of anti graffiti coatings on stone natural vs accelerated weathering
K Factor: 2.57
Fields: weathering, visual arts, graffiti, durability, commons, art

Title: hybrid numerical and experimental performance assessment of structural thermal bridge retrofits
K Factor: 2.53
Fields: urdu, tamil, serbian, portuguese, malayalam, macedonian, latvian, indonesian, history, ancient history

Subject: environmental_sciences
==========

Title: environmental impact of switching from the synthetic glucocorticoid prednisolone to the natural alkaloid berberine
K Factor: 2.35
Fields: ranging, prednisolone, pharmacology, glucocorticoid, environmental science, defined daily dose, berberine, alkaloid

Title: multiple detection of zoonotic variegated squirrel bornavirus 1 rna in different squirrel species suggests a possible unknown origin for the virus
K Factor: 2.35
Fields: zoology, zoological garden, variegated squirrel, tamiops swinhoei, subfamily, sciurus granatensis, sciurus, callosciurus, callosciurinae, biology

Title: boerhaave s mineral chemistry and its influence on eighteenth century pharmacy in the netherlands and england
K Factor: 2.35
Fields: philosophy, pharmacy, pharmacist, nothing, nomenclature, literature, chemistry, chemical transformation, alternative medicine, academic medicine

Title: listening to earthworms burrowing and roots growing acoustic signatures of soil biological activity
K Factor: 2.33
Fields: soil structure, soil science, plants root, environmental science, ecosystem ecology, earthworm, burrow, biophysical processes, agroecology

Title: evidence for the breakdown of an angkorian hydraulic system and its historical implications for understanding the khmer empire
K Factor: 2.31
Fields: weir, levee, juncture, hydraulic machinery, history, empire, civil engineering, archaeology

Subject: humanities
==========
Title: literary evidence for taro in the ancient mediterranean a chronology of names and uses in a multilingual world
K Factor: 1.68
Fields: nelumbo, medicinal plants, lotus, latin literature, history, colocasia esculenta, colocasia, arum, ancient literature, ancient history

Title: the dangers and promises of comparative history of science
K Factor: 1.62
Fields: spanish civil war, performance art, nazism, history of science and technology, history of science, history, encyclopedia, comparative history, classics, civilization, china

Title: conceptions of self determination in fourth tenth century muslim theology al bāqillānī s theory of human acts in its historical context
K Factor: 1.58
Fields: voluntariness, theology, self determination, philosophy, moral responsibility, extant taxon

Title: a revision of sanpasaurus yaoi young 1944 from the early jurassic of china and its relevance to the early evolution of sauropoda dinosauria
K Factor: 1.56
Fields: zoology, vulcanodontidae, ungual, sauropoda, sanpasaurus, paleontology, ornithopod, eusauropoda, biology, basal, autapomorphy

Title: digitally reconstructing the great parchment book 3d recovery of fire damaged historical documents
K Factor: 1.56
Fields: visual arts, population, politics, parchment, irish, fragile state, engineering, digital history, digital edition, custodians

Subject: maths_computing_ee
==========
Title: alteration of marble stones by red discoloration phenomena
K Factor: 2.60
Fields: mining engineering, art

Title: stop beating the donkey a fresh interpretation of conditional donkey sentences
K Factor: 2.48
Fields: situation semantics, philosophy, generalized quantifier, experimental data, donkey, algorithm

Title: enabling rootless linux containers in multi user environments the udocker tool
K Factor: 2.43
Fields: operating system, multi user, mathematics, functional testing

Title: zakat accounting metaphor and accounting treatment for business organization
K Factor: 2.42
Fields: positive accounting, mark to market accounting, management accounting, generally accepted accounting principles, financial accounting, entity concept, business, accounting information system, accounting identity, accounting

Title: forecasting day ahead electricity prices in europe the importance of considering market integration
K Factor: 2.41
Fields: symmetric mean absolute percentage error, market integration, functional analysis, financial economics, feature selection, electricity price forecasting, electricity market, economics, bayesian optimization, artificial neural network

Subject: medical_sciences
==========
Title: educational expansion and inequalities in mortality a fixed effects analysis using longitudinal data from 18 european populations

K Factor: 2.65
Fields: demographic economics, ceteris paribus

Title: thyroid hormone action and disruption in the brain xenopus as a model to study disruption of thyroid hormone availability on early brain development
K Factor: 2.65
Fields: molecular biology, history

Title: dueling biological and social contagions
K Factor: 2.65
Fields: political science, alternative medicine

Title: chromatographic methods for nannochloropsis gaditana microalgae extracts profiling
K Factor: 2.53
Fields: nannochloropsis gaditana, chromatography, art

Title: perverse conservatism a lacanian interpretation of russia s turn to traditional values
K Factor: 2.51
Fields: traditional values, subjectification, sociology, social science, psychoanalytic theory, perversion, persecution, fetishism, domestic policy, conservatism

Subject: physics
==========
Title: multi messenger observations of a binary neutron star
K Factor: 3.66
Fields: spitzer space telescope, sociology, observatory, meter, engineering physics

Title: generalized laws of thermodynamics in the presence of correlations
K Factor: 3.55
Fields: thermodynamics, thermal management of electronic devices and systems, refrigeration, laws of thermodynamics, information flow, helmholtz free energy, genetics, erasure, conditional entropy, biology

Title: planetary protection of outer solar system bodies
K Factor: 3.46
Fields: solar system, planetary protection, jupiter, icy moon, galilean moons, environmental science, enceladus, astrobiology

Title: tandem accelerators in romania multi tools for science education and technology
K Factor: 3.45
Fields: tandem, simulation, science education, revenue, proton therapy, particle accelerator, ion implantation, engineering, counterfeit, computer engineering

Title: design and modeling of an additive manufactured thin shell for x ray astronomy
K Factor: 3.38
Fields: x ray telescope, x ray astronomy, space technology, sextant, polishing, mechanical

engineering, engineering, electronic engineering, 3d printing

Subject: social_sciences
==========
Title: the norm4building database a tool for radiological assessment when using by products in building materials
K Factor: 2.43
Fields: radiological weapon, materials science, database

Title: health examination surveys and human biomonitoring the added value of combined studies
K Factor: 2.43
Fields: sociology, optometry, functional testing, added value

Title: effect of fluid viscosity on noise of bileaflet prosthetic heart valve
K Factor: 2.36
Fields: prosthetic heart, physics, nuclear chemistry

Title: biometric measurements in the crystalline lens applications in cataract surgery
K Factor: 2.36
Fields: philosophy, ophthalmology

Title: the island is not a story in itself apartheid s world literature
K Factor: 2.32
Fields: world literature, south african literature, sociology, moral conditions, international community, hieroglyph, anthropology

Pilot 8: Linkages and Knowledge Exchange Indicators

Abstract:
This pilot investigates the linkages between research outputs of Research and Innovation (R&I) projects in the Horizon 2020 programme of the European Commission. Complex network analysis techniques are applied to model the linkages between research grants, organisations, researchers, and publications and to analyse correlations between allocated funding and collaborations and knowledge exchange in Europe. A set of indicators for quantifying these linkages are discussed and the suggested methodology is applied on datasets representing two R&I ecosystems.

1 Introduction

1.1 Background/context

Scientific knowledge is almost always built cumulatively on top of previous discoveries (Scotchmer, 1991). Within the context of research and innovation (R&I) ecosystems, knowledge is transferred in many ways between the entities of an ecosystem: through academic publications and patents, face-to-face discussions between researchers, collaborations on an organisational level when working on a shared research project. Thus, it is important to measure the knowledge flow between various elements of the ecosystem that enables such scientific progress. While the current R&I indicators in the European Union mostly measure the quantity of R&I linkages (Hollanders and Es-Sadki, 2018), during the EURITO knowledge stakeholder workshop (KSW), participant policy makers expressed the need to know not only the amounts, but also the nature of these linkages to gain more insights about the impact of the funded projects on the knowledge flow in Europe.

Horizon 2020 is Europe's largest research and innovation funding program with the overall budget of 77 billion euros ("Horizon 2020 Work Programme from 2018 to 2020", 2019). Understanding how such a vast amount of funding allocated to the R&I projects fosters collaboration and knowledge exchange between research participants is, therefore, critical to evaluate the directions of future funding efforts. Thus, viewing a R&I ecosystem through networks is necessary to build the understanding of the overall structure and behaviour of the ecosystem, as collaborations between ecosystem members lead to creation of unique combination of competencies, which boosts innovation and scientific progress compared to a single organisation, discipline or location (Kang and Hwang, 2016). For instance, Ortega and Aguillo (2010) showed that countries that have central positions in the network also have shown higher R&D capacity.

In essence, EU has always supported creation of R&I networks through Framework Program (FP) and Horizon 2020 programmes. Indeed, both programmes promote networks through knowledge exchange and collaboration within consortiums that are normally created during the implementation of the program projects (Kang and Hwang, 2016). However, while there are currently available indicators that assess the phenomena of collaboration and knowledge exchange between the European Commission projects, we propose to further strengthen them so that they could not only report the aggregated amount of such linkages, but also allow policy makers to zoom into details about the nature of these linkages. Moving from tabular data view about linkages in R&I ecosystem to network view, where these connections and collaboration trends become visible would be a step towards transparency and betterment of research funding.

1.1.1 Opportunity

In this pilot we attempt to model and analyse the nature of collaborations and knowledge exchange in R&I ecosystems of two different scales and with two different datasets: European Commission's Horizon 2020 program and a research funding agency in Denmark (further referred to as FAD). For both datasets, we employ complex network analysis techniques to quantify the nature of linkages between projects, organisations, researchers and research outputs. Complex network analysis enables one to explore the structural characteristics of R&I ecosystems. By analysing complex network measures, such as node degree, betweenness and eigenvector centralities, we can determine entities in the ecosystems that are central, connected and serve as gateways between the other entities.

A study done previously in European Commission (Science-Metrix et al., 2015), performs social network analysis on FP7 data, however only uses citation links between publications, while the aim of our pilot is to capture connections beyond citations, e.g. links to grants, organisations, researchers and respective research outputs. Furthermore, our focus is on expressing these linkages through a set of quantitative indicators.

Another similar study was conducted using National Science Foundation (NSF) data in USA, where three types of networks were constructed to assess the knowledge exchange related to the research projects: primary investigators, organisations and countries (Kardes et al., 2014). Though, besides a different dataset, key conceptual differences of our pilot is that we analyse all research entities within a single network and we add information not only about actors that perform research (i.e. organisations or researchers), but also about the related research outputs (e.g. publications).

1.1.2 Application domain

This pilot uses datasets connected to the reporting of R&I project funding. While the dataset from FAD contains projects from the specific domain, Horizon-2020 dataset covers projects across all R&I topics.

1.1.3 Flexibility of application domain

The indicators resulting from the current pilot are designed in a way so that they can be measured independently of the application domain, as long as research outputs are linked together. In essence, Horizon-2020 dataset has combined different domains ranging from healthcare to aerospace technology.

1.1.4 Assessment of 'periphery to core of R&I policy' capacity of proposed pilot

Currently, there is a lack of indicators in R&I policy, that would describe the knowledge flow and R&I linkages between heterogeneous research entities (projects, research outputs, organisations). The semantically closest indicators from the European Innovation Scoreboard are "International scientific co-publications" under the "Attractive research systems" section and all indicators under the "Linkages" section: "Innovative SMEs collaborating with others", "Public-private co-publications" and "Private co-funding of public R&D expenditures".

While these indicators show the amount of linkages and collaborations between research entities, there is a need for more detailed indicators that would as well illustrate not only the quantity, but also the nature of these linkages.

1.1.5 Stakeholder engagement summary

The interest of this type of assessment of funding impact was expressed by the executives of the FAD. FAD provided us with the internal project data, feedback for the preliminary findings, as well as suggestions on the further directions of the pilot.

Moreover, rather positive feedback was received at the KSW. One of the suggestions from the policy experts was to consider that knowledge flow occurs not through one, but through multiple channels. Thus, indicators proposed in this pilot are based on complex network measures that allow to find those channels – research entities that link together various subnetworks of the R&I ecosystem.

1.2 Relevance to RITO criteria

1.2.1 Relevant

The high **relevancy** of developing an indicator that would capture collaborations and knowledge exchange across R&I projects was highlighted several times at the KSW. In particular, policy makers and executives of funding agencies expressed the need for alternative ways to evaluate the impact of the publicly funded research projects.

1.2.2 Inclusive

First, variety of subjects across R&I projects in two major funding schemes considered in the Horizon-2020 dataset contributes to the sectoral **inclusiveness** of the indicator.

Second, Horizon-2020 is the largest R&I program in Europe and covers all countries in EU, which contributes to the **geographical inclusiveness** within the European Union.

1.2.3 Timely

The indicators are **timely**, as the records in the Horizon-2020 dataset are updated on an ongoing basis and allow various levels of granularity.

1.2.4 Trusted

The indicators are considered **trusted**, as they originate from European Open Data Portal, which is officially curated data source, supported by European Commission.

1.2.5 Open

The indicators and the methodology are publicly available and are practically identical for both datasets. While FAD dataset can not be disclosed at a more detailed level and is used only for internal validation with stakeholders, the Horizon-2020 data is **open** and can be used for the further analyses.

1.3 Research/policy questions

The pilot attempts to enable the policy makers to answer the following questions:

- What is the nature of the knowledge flow resulted from the publicly funded R&I projects?
- What are the most central organisations in the funding programme?

- What research projects create the most “bridges” between various organisations, researchers and research outputs?

- How does the collaboration network and knowledge exchange enabled by research projects correlate with the funding they have received?

2 Methodology

2.1 Data sources

Complex network analysis was performed using two major datasets of projects funded by a funding organization in Denmark and European Commission’s Horizon 2020 program from European Open Data Portal (https://data.europa.eu/euodp/en/data). This section discusses the initial steps of data collection and preprocessing before applying complex network analysis.

Horizon 2020 and OpenAire datasets
This dataset was constructed from two major data sources:

- European Open Data Portal (https://data.europa.eu/euodp/en/data), which contains data on projects funded by European Commission’s Horizon 2020 program (covers years from 2014 onwards).
- OpenAire portal (https://www.openaire.eu), which contains data about open-access publications that were funded by European Commission’s Horizon 2020 program.

First, 20 878 projects funded by the European Commission under Horizon 2020 program and 27 510 connected organisations and 24 208 involved researchers linked to these projects and organisations were retrieved from Horizon2020 and processed using Jupyter Notebook and Python. Finally, we linked this list of projects to 272 470 publications retrieved from the OpenAire dataset.

During the preprocessing, publications without the information about the date of acceptance and without the DOI identifier were removed. The research entities that were not linked to any other research entities, i.e. with the degree equal to zero, were also not considered in the analysis Finally, a list of projects, publications, organisations and researchers was converted into nodelist and edgelist structure, suitable for the further construction of a network. The final number of nodes is 105 567 research entities with 150 615 links (edges) in between them.

Finally, the resulted dataset was converted to network representation using Jupyter Notebook. After we have performed all the data preprocessing and analysis in Jupyter Notebook, the Graphml file was generated and visually validated via Gephi – a visual network analysis tool.

FAD dataset
From the FAD dataset, which was provided by the data owners in the Microsoft Excel workbook format, 68386 entities of various types such as publications, patents, datasets, spinouts, awards, artistic products and other types research outputs were extracted (17 types in total).

All these outputs were preprocessed so that they could be linked to specific grant identifiers. Besides, each output contained a time information – e.g. a year when the grant was received. Since grants did not have an explicit year information, the years were assigned according to a date when the grant was

awarded. If an output had an empty year, the year for this output was assigned equal to a year of the grant connected to this output.

After the removal of the duplicates (due to the overlapped reporting across several reporting periods) and removing the grants without produced outputs, the dataset contained 22 976 records.

2.2 Methods

This section describes the common methodology that was used to perform the complex network analysis for both datasets. First, connections between all the research related entities (projects, organisations, publications, researchers), were extracted from the datasets. For instance, a connection between a publication and a project was made, if a publication record was tied to a particular research project identifier. Records that contain research entities were converted into a node list of the network, while the connections were treated as edges between the nodes and form an edge list of the network. Having both node list and edge list allows one to construct the whole network and perform the subsequent network analysis.

In the next step, for each node in the network we have calculated the following network measures using Python iGraph library's built-in algorithms (Kardes et al., 2014):

Degree. Degree of a node denotes the number of relations that this node has with other nodes. For instance, in the public policy context, a high number of degrees of a node that represents a project means that this project has various connections to other research entities: related publications, involved organisations or researchers that worked on this project.

Betweenness centrality. Betweenness centrality of a node measures the number of times shortest paths between all the nodes in the network pass through that particular node. In essence, betweenness centrality estimates the node's ability to serve as a "bridge" that connects different network parts.

Eigenvector centrality: Eigenvector centrality of a node estimates how important is that particular node within the whole network. This importance is determined by how central are the nodes that connect to that node. Eigenvector centrality gives less weight to the number of connections to the node, but prioritises the quality of connections – a node might have a small number of connections, but these connections might have high centrality values.

During the calculation both betweenness and eigenvector centralities were normalised according to the maximum values in order to alleviate the comparison of different research entities.

Centrality – Funding ratio (CFR score). It is expected that research projects that receive large amounts of funding would have large number of connections, and therefore, larger centrality values. To account for this, we calculate the ratio between centrality measures of research projects and the amount of funding they receive.

This measure allows to understand whether the amount of funding for research projects is correlated with their "connectedness". CFR score is calculated by simply dividing the centrality (either betweenness or eigenvector) value of a node that represents a project by the amount of funding that this particular project have received.

$$CFR(n_i) = \frac{C_b(n_i)}{F(n_i)}$$

where

n_i = node of a network

C_b (n_i) = betweenness or eigenvector centrality value for this node,

F (n_i) = the amount of funding allocated to this node ("ecMaxContribution" column in the Horizon2020 dataset)

2.3 Documentation

Code and generated data for this pilot are available at the link below:
https://github.com/EURITO/wp2pilots/tree/master/Pilot8

The code is in Jupyter Notebook format and broken down into five main steps:

1. Data retrieval from datasets
2. Parsing retrieved data
3. Transforming data to network format
4. Complex network analysis
5. Output to human-readable format (Gephi compatible *.graphml files)

3 Results

3.1 Horizon-2020 dataset

Betweenness Centrality.

As seen in Figure 1, betweenness centrality measure is able to show how "connected" are the research entities in the overall R&I ecosystem and how well the generated knowledge propagates across the network. For instance, research organisations such as Centre National de la Recherche Scientifique (CNRS, France), which is the largest fundamental science agency in Europe (Butler, 2008) or Fraunhofer (Germany), which is one of the largest applied research organisations in Europe (Fraunhofer, 2019) are represented by large nodes (betweenness centrality values > 0.9).

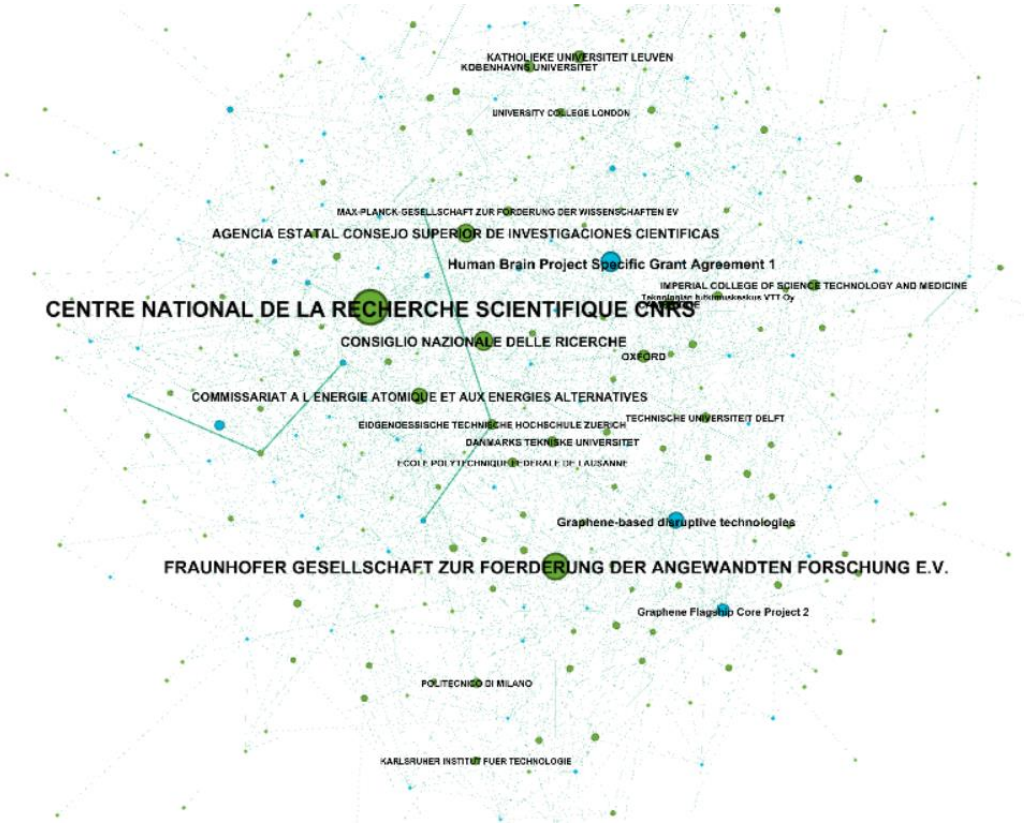


Figure 1. Betweenness centrality across Horizon2020 projects (entities, where betweenness centrality > 0.019 are shown).
Larger node sizes represent larger betweenneess centrality values, while colors represent the type of research entity: turqouse– projects, green – organisations

Further, Table 1 shows the top 20 research organisations sorted by their betweenness centrality values. As seen from the table, these organisations are, indeed, constitute the list of the largest research institutions in their respective countries, and which are more likely to participate in Horizon 2020 research projects.

Country	**Organisation**	**Betwenness Centrality**	**Degree**
FR	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	1	1078
DE	FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.	0.729927	714
IT	CONSIGLIO NAZIONALE DELLE RICERCHE	0.485563	538
ES	AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	0.46064	542
FR	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	0.381003	453
UK	THE CHANCELLOR, MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD	0.269359	482
BE	KATHOLIEKE UNIVERSITEIT LEUVEN	0.265989	490
UK	THE CHANCELLOR MASTERS AND SCHOLARS OF THE UNIVERSITY OF CAMBRIDGE	0.254045	497
UK	IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE	0.231561	387

DK	KOBENHAVNS UNIVERSITET	0.228866	481
DK	DANMARKS TEKNISKE UNIVERSITET	0.218097	359
CH	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	0.203037	336
NL	TECHNISCHE UNIVERSITEIT DELFT	0.196072	363
CH	ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE	0.181722	384
DE	MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV	0.178954	376
IT	POLITECNICO DI MILANO	0.177736	278
UK	THE UNIVERSITY OF EDINBURGH	0.163546	297
UK	UNIVERSITY COLLEGE LONDON	0.163239	374
FI	Teknologian tutkimuskeskus VTT Oy	0.148531	247
DE	KARLSRUHER INSTITUT FUER TECHNOLOGIE	0.133289	210

Table 1. Top 20 organisations according to the betweenness centrality in the Horizon2020 program

Accordingly, betweenness centrality can indicate the most central research projects as well. For instance, two of the largest research projects in the European Commission, Graphene projects 1 & 2 ("Graphene Flagship", 2019) when combined span over 150 organisations and 23 countries. They were both highlighted with larger sized nodes in Figure 1. In another example, the Human Brain Project is a ten-year project that started in 2013 and spans more than 100 research organisations (https://www.humanbrainproject.eu/en/about/overview/). Human Brain Project scored higher than 0.5 on the betweenness centrality score and has 416 connections. Though, only 4 projects achieve betweenness centrality values larger than 0.1

Project acronym	**Project title**	**Betweenness Centrality**	**Degree**	**EC contribution**	**CFR score**
HBP SGA1	Human Brain Project Specific Grant Agreement 1	0.506913	416	89000000	5.70E-09
GrapheneCore1	Graphene-based disruptive technologies	0.391452	330	89000000	4.40E-09
GrapheneCore2	Graphene Flagship Core Project 2	0.285184	146	88000000	3.24E-09
EUROfusion	Implementation of activities described in the Roadmap to Fusion during Horizon 2020 through a Joint programme of the members of the EUROfusion consortium	0.195325	928	44080000	4.43E-10
InvisiblesPlus	InvisiblesPlus	0.080825	197	2070000	3.90E-08
AIDA-2020	Advanced European Infrastructures for Detectors at Accelerators	0.080059	137	10000000	8.01E-09
ELIXIR-EXCELERATE	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life-sciences.	0.071614	113	19051482	3.76E-09

ECOPOTENTIAL	ECOPOTENTIAL: IMPROVING FUTURE ECOSYSTEM BENEFITS THROUGH EARTH OBSERVATIONS	0.061253	117	14874340	4.12E-09
ENSAR2	European Nuclear Science and Application Research 2	0.058099	171	10000000	5.81E-09
OpenAIRE2020	Open Access Infrastructure for Research in Europe 2020	0.048343	104	13000000	3.72E-09
OPTICON	Optical Infrared Coordination Network for Astronomy	0.048011	35	10000000	4.80E-09
ACTRIS-2	Aerosols, Clouds, and Trace gases Research InfraStructure	0.047118	195	9541194	4.94E-09
EOSC-hub	Integrating and managing services for the European Open Science Cloud	0.046636	77	30000000	1.55E-09
Productive4.0	Electronics and ICT as enabler for digital industry and optimized supply chain management covering the entire product lifecycle	0.044561	111	26033148	1.71E-09
ZIKAlliance	A global alliance for Zika virus control and prevention	0.044051	86	11964209	3.68E-09
SHARE-DEV3	Achieving world-class standards in all SHARE countries	0.043594	215	5493328	7.94E-09
TBVAC2020	TBVAC2020; Advancing novel and promising TB vaccine candidates from discovery to preclinical and early clinical development	0.043233	98	18200000	2.38E-09
LASERLAB-EUROPE	The Integrated Initiative of European Laser Research Infrastructures	0.041958	161	10000000	4.20E-09
BIORIMA	BIOmaterial RIsk MAnagement	0.041359	41	7999981	5.17E-09
RItrain	Reseach Infrastructures Training Programme	0.041272	13	1995634	2.07E-08

Table 2. Top 20 research projects according to the betweenness centrality in Horizon2020 program

While these larger projects are relatively transparent, policy makers might be interested to explore how these betweenness centrality values correlate with the funding amount that these projects have received or allocated to receive from the European Commission. To analyse this ratio, we have calculated the CFR score defined in Section 2.2 across the network to find projects with the highest betweenness centrality to received funding ratio. Illustrated in Table 3, this list shows examples of top 20 projects that were able efficiently use allocated funding to achieve intensive knowledge flow through reaching out to a large number of organisations and researchers or producing large number of publications.

Project acronym	**Project title**	**Betweenness Centrality**	**Degree**	**EC contribution**	**CFR score**

StronGrHEP	Strong Gravity and High-Energy Physics	0.024751	82	288000	8.59E-08
BITNET-INNOSUP	Improvement of Innovation Management Capacity of SMEs from Bosnia and Herzegovina (FBH) through the Enterprise Europe Network	0.00082	7	13045	6.29E-08
SILKENE	SILKENE: Bionic silk with graphene or other nanomaterials spun by silkworms	0.008971	45	149944	5.98E-08
Business INN Moldova	Business INN Moldova	0.000816	5	14225	5.73E-08
EuroStem Cell	European Consortium for Communicating Stem Cell Research	0.034354	31	600000	5.73E-08
NonMinimalHiggs	Non Minimal Higgs	0.015345	66	301500	5.09E-08
PROTINUS	PROviding new insighT into INteractions between soil fUnctions and Structure	0.008834	18	175500	5.03E-08
QUANTUM DYNAMICS	New Geometry of Quantum Dynamics	0.012901	26	288000	4.48E-08
SYSMICS	Syntax Meets Semantics: Methods, Interactions, and Connections in Substructural logics.	0.021104	36	504000	4.19E-08
Invisibles Plus	InvisiblesPlus	0.080825	197	2070000	3.90E-08
WEST-MED Innovation	WEST-MED Innovation Services	0.000612	4	15746	3.89E-08
PEER FOR EXCELLENCE	Peer learning on ways to enhance good practices in SME innovation support using the Seal ofExcellence	0.001852	5	50000	3.70E-08
BinCosmos	The impact of Massive Binary Stars through Cosmic Times	0.006117	31	165598	3.69E-08
GEAGAM	Geophysical Exploration using Advanced GAlerkin Methods	0.021283	103	580500	3.67E-08
INNO RO 4 EUROPE	Enhancing economic impact in SME's in Romania by building innovation management capacity	0.000408	4	11168	3.65E-08
NEXT-3D	Next generation of 3D multifunctional materials and coatings for biomedical applications	0.007014	17	193500	3.63E-08
SME Growth	AcceleGreat - SME Growth	0.001722	3	50000	3.44E-08
FourCmodelling	Conflict, Competition, Cooperation and Complexity: Using	0.007276	27	216000	3.37E-08

	Evolutionary Game Theory to model realistic populations				
EUNORS	Enhancing innovation management capacity of SMEs in Republic of Srpska	0.000614	5	20052	3.06E-08
Diaspora Link	DiasporaLink	0.016339	29	571500	2.86E-08

Table 3. Top 20 research projects according to the CFR score

When sorting projects by CFR score, we discovered that the majority of the top 20 research projects were on the topic of improving innovation management of SME's. This can be explained by the fact that these smaller grants require less capital costs, compared to more technical projects that require purchase of equipment, materials, etc, and, at the same time required participation of a large number of organisations.

Eigenvector centrality.

Measuring eigenvector centrality for organisations produced similar results to the betweenness centrality, which can be observed in Table 4. While 19 out of top 20 research institutions remained the same compared to Table 1 (except the University of Manchester which entered the list instead of the University of Edinburgh), Karlsruhe Institute of Technology (KIT) have moved 13 positions up in the list. Network interpretation of this event might indicate that even though KIT has less connections to compared to the institutions that were above previously in the betweenness centrality list, these connections to research projects themselves possess higher centrality values.

Country	**Organisation**	**Eigenvector Centrality**	**Degree**
FR	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	1	1078
DE	FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.	0.40988537	714
FR	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	0.307901532	453
IT	CONSIGLIO NAZIONALE DELLE RICERCHE	0.282525133	538
ES	AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	0.267585447	542
CH	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	0.150242941	336
DE	KARLSRUHER INSTITUT FUER TECHNOLOGIE	0.136653946	210
DK	DANMARKS TEKNISKE UNIVERSITET	0.135460022	359
UK	THE CHANCELLOR, MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD	0.135431741	482
FI	Teknologian tutkimuskeskus VTT Oy	0.128306027	247
CH	ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE	0.127002075	384
UK	THE CHANCELLOR MASTERS AND SCHOLARS OF THE UNIVERSITY OF CAMBRIDGE	0.124327178	497
BE	KATHOLIEKE UNIVERSITEIT LEUVEN	0.118489277	490
NL	TECHNISCHE UNIVERSITEIT DELFT	0.117380252	363
DE	MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV	0.116162559	376

UK	IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE	0.114547992	387
DE	DEUTSCHES ZENTRUM FUER LUFT - UND RAUMFAHRT EV	0.104555281	276
UK	THE UNIVERSITY OF MANCHESTER	0.102828764	250
DK	KOBENHAVNS UNIVERSITET	0.100103741	481
IT	POLITECNICO DI MILANO	0.099288381	278

Table 4. Top 20 organisations according to the eigenvector centrality in the Horizon2020 program

Calculation of the eigenvector centrality for research projects showed more variance compared to the betweenness centrality. Table 5 shows that even though top 4 projects remained in the same area of the list, only 5 more projects remained in the top 20.

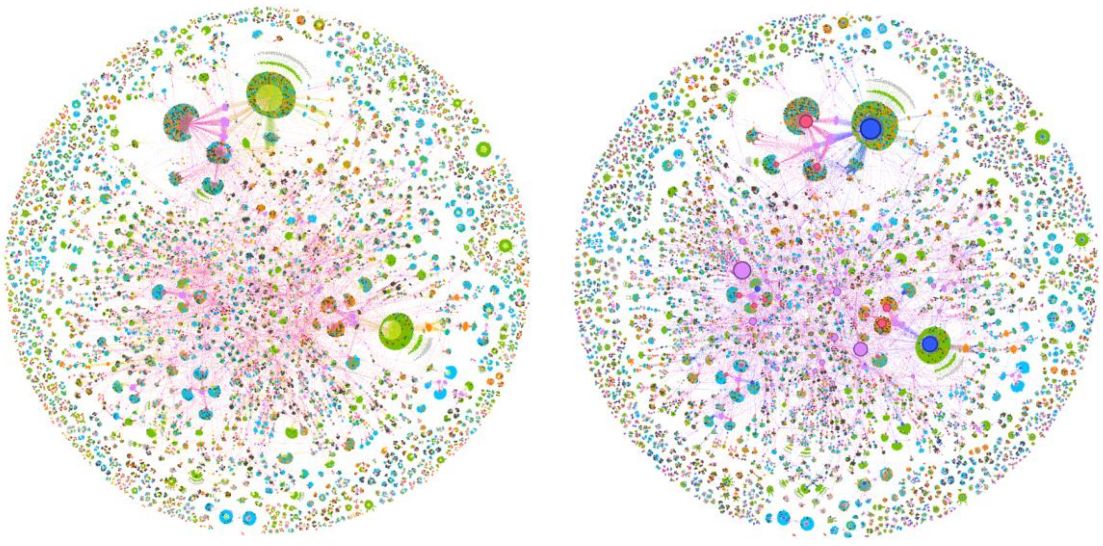
Project acronym	**Project title**	**Eigenvector Centrality**	**Degree**	**EC contribution**	**CFR score**
Graphene Core1	Graphene-based disruptive technologies	0.235601758	330	89000000	2.65E-09
HBP SGA1	Human Brain Project Specific Grant Agreement 1	0.215399962	416	89000000	2.42E-09
Graphene Core2	Graphene Flagship Core Project 2	0.188515028	146	88000000	2.14E-09
EUROfusion	Implementation of activities described in the Roadmap to Fusion during Horizon 2020 through a Joint programme of the members of the EUROfusion consortium	0.117859256	928	4.41E+08	2.67E-10
EOSCpilot	The European Open Science Cloud for Research Pilot Project.	0.103349413	49	9953068	1.04E-08
EOSC-hub	Integrating and managing services for the European Open Science Cloud	0.100071963	77	30000000	3.34E-09
AIDA-2020	Advanced European Infrastructures for Detectors at Accelerators	0.095703478	137	10000000	9.57E-09
ECOPOTENTIAL	ECOPOTENTIAL: IMPROVING FUTURE ECOSYSTEM BENEFITS THROUGH EARTH OBSERVATIONS	0.085304458	117	14874340	5.74E-09
ELIXIR-EXCELERATE	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life-sciences.	0.082057649	113	19051482	4.31E-09
ARIES	Accelerator Research and Innovation for European Science and Society	0.080511044	58	10000000	8.05E-09
EPOS IP	EPOS Implementation Phase	0.076768777	70	18374344	4.18E-09
ENVRI PLUS	Environmental Research Infrastructures Providing Shared Solutions for Science and Society	0.075263961	56	14683534	5.13E-09

INSHIP	Integrating National Research Agendas on Solar Heat for Industrial Processes	0.074611484	30	2498661	2.99E-08
ACTPHAST 4.0	ACceleraTing PHotonics innovAtion for SME's: a one STop-shop-incubator	0.072371159	25	9999946	7.24E-09
ERA4CS	European Research Area for Climate Services	0.072353241	46	25000000	2.89E-09
EGI-Engage	Engaging the EGI Community towards an Open Science Commons	0.071327381	48	8000000	8.92E-09
ACTPHAST 4R	Accelerating Photonics Deployment via one Stop Shop Advanced Technology Access for Researchers	0.070908693	24	6000000	1.18E-08
MyOcean FO	Pre-Operational Marine Service Continuity in Transition towards Copernicus	0.070766427	59	6000000	1.18E-08
INDIGO-DataCloud	INtegrating Distributed data Infrastructures for Global ExplOitation	0.070073723	34	11138114	6.29E-09
ENSAR2	European Nuclear Science and Application Research 2	0.070058112	171	10000000	7.01E-09

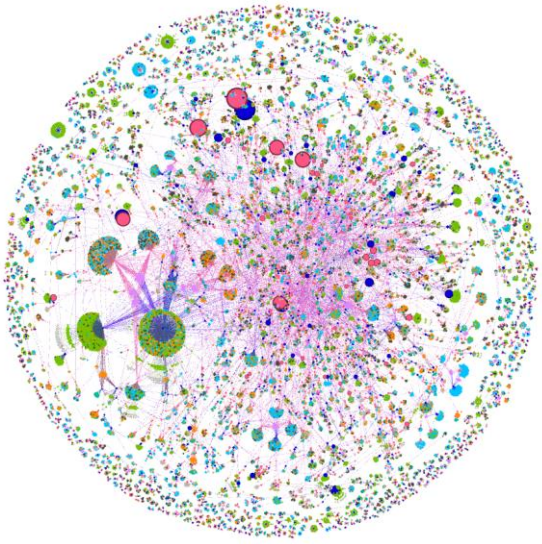
Table 5. Top 20 research projects according to the eigenvector centrality in the Horizon2020 program. Projects highlighted by bold font indicate the research projects that remain in top 20 compared to top 20 by betweenness centrality measure in Table 2.

3.2 FAD dataset

Due to the data sharing agreement between DTU and FAD, the exact data entities and respective analysis cannot be included in this report. However, parts of the methodology in Section 2 were developed based on the obtained feedback on the preliminary results of the analysis. In this report, we include the visualization of the analysis to provide an overview of the R&I ecosystem of this particular funding organisation.



A) B)



C)

Figure 2 – A) FAD dataset network (nodes of equal sizes) B) node size are set to represent values of betweenness centrality for each node C) node sizes represent larger CFR values
Colors represent types of nodes: blue – institutional projects, red – research grants, green – publications

4 Discussion and Conclusions

In this pilot, we have applied network analysis techniques to two datasets that contain data about research projects under the European Commission's Horizon 2020 program, participant organisations, researchers and respective academic publications linked to those projects. While preliminary, the results have demonstrated that it is possible to analyse such data in a novel fashion and obtain potentially applicable insights. For instance, when analysing complex network measures for research projects, we were able to verify that the largest organisations and projects of the Horizon 2020 program will be the most central ones as well. In a similar fashion the largest and the most central research projects in Europe were identified.

Overall, given the available data, the proposed approach allows to measure collaborations and knowledge exchange within a particular funding program. Such information is valuable to explore what projects are highly connected to other elements in the R&I ecosystem, as it allows to extend the understanding of policy makers about the performance of the projects beyond standard characteristics, e.g. the number of produced publications or the amount of R&I linkages.

Using the outlined approach, policy makers can answer the policy questions regarding the most central organisations and projects in the R&I ecosystem in Europe. By zooming into connections of each of these research entities, the nature of the relationships between these central research entities could be analysed. Moreover, the proposed CFR score attempts to find a ratio between the "connectivity" of the research projects and the amount of funding that these projects receive.

When extended, this approach allows policy makers to examine data in different levels of granularity. In the simplest form, network measures analysed above could be calculated not only for organisations, but likewise for both higher and lower levels of detail - e.g. per country or per individual researchers.

In addition, we have analysed data sources of two different types of funding programs: within one funding agency and within the whole funding program. Thus, the methodology outlined in this pilot allows the further comparison of resulted the network structures for R&I ecosystems of different scales.

4.1 Validation and ongoing stakeholder engagement

We plan to continue obtain feedback from the policy makers in Europe and validate the findings outlined here through future research and knowledge dissemination events.

4.2 Limitations

The main limitation of the pilot for now is the chosen modelling scheme of the linkages, as only direct and evident connections between research grants, organisations, researchers and publications are considered. This could be extended by finding more connections within the research entities (e.g. through Natural Language Processing techniques and topic modelling).

The OpenAire dataset has following limitations:

- Software records are provided under selective access API, which limits the total number of extractable records to 10000. While it is not a problem for this pilot, as there are 4231 records of EC funded software, it may introduce limitations in future, as dataset grows. Bulk access API does not provide an endpoint for software records.
- OAI-PMH standard does not support incremental harvesting as of December 2018, so the whole dataset has to be retrieved from scratch on every update.
- We have not removed some obvious outliers, as we included all the records, which had the necessary fields. For instance, large constellation on top right of Figure 2 stems from a single publication DOI, which in truth is a funding scheme DOI, instead of an actual publication DOI.

Further development of this pilot would include the following:

- Incorporation of additional datasets (e.g. research datasets, patents)
- Analysis of call topics of the Horizon 2020 program
- Development of interactive visualisation
- Temporal breakdown and dynamic network analysis

4.3 Considerations for scaling up

4.3.1 Complementarities with other pilots

This pilot could be complemented with Pilot 6 “Advanced R&I funding analytics”, as they both aim at similar issues: team collaboration, funding impact, diversity in projects and dissemination of research outputs.

4.3.2 Tools and data sources

In the pilot, we have experimented with two data infrastructure approaches (Table 6). While for FAD data we have used an approach with more exploratory and manual steps, the Horizon-2020 dataset was preprocessed and analysed using a more integrated approach, without much of a manual work. This allows to extend created infrastructure for other similar datasets in future.

Framework	Data collection	Cleaning	Network transformation	Complex network analysis	Presentation
Exploratory	Microsoft Excel	Microsoft Excel, Google OpenRefine	Tab2Net	Gephi, Excel	Tabular, Gephi visualisation
Integrated	Python/Jupyter Notebook		NetworkX/iGraph within Python/Jupyter Notebook		Tabular, Gephi visualisation

Table 6. Comparison of tools used in the pilot

5 References

1. Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." Journal of Economic Perspectives, 5 (1): 29-41.
2. Butler, D. (2008), “France’s research agency splits up”, *Nature*, Nature Publishing Group, Vol. 453 No. 7195, p. 573, https://doi.org/10.1038/453573a.
3. Fraunhofer. (2019), “About Fraunhofer: Facts and Figures”. Retrieved from https://www.fraunhofer.de/en/about-fraunhofer/profile-structure/facts-and-figures.html
4. “Graphene Flagship”. (2019). Retrieved from http://graphene-flagship.eu/
5. Hollanders, H., Es-Sadki, N. (2018), *European Innovation Scoreboard 2018*, https://doi.org/10.1007/s13398-014-0173-7.2.
6. “Horizon 2020 Work Programme from 2018 to 2020”. (2019). Retrieved from https://ec.europa.eu/programmes/horizon2020/en/news/horizon-2020-work-programme-2018-2020
7. Kardes, H., Sevincer, A., Gunes, M.H. and Yuksel, M. (2014), “Complex Network Analysis of Research Funding: A Case Study of NSF Grants”, pp. 163–187, https://doi.org/10.1007/978-3-319-05912-9_8.
8. Science-Metrix, Frauenhofer ISI and OXFORD Research. (2015), *Study on the Network Analysis of the 7th Framework Programme Participation: Methodological Annex*.
9. Ortega, J.L. & Aguillo, I.F. Scientometrics (2010) 84: 835. https://doi.org/10.1007/s11192-010-0212-x

Pilot 7: Inclusive Innovation

How inclusive is Europe's digital economy? An analysis of gender and ethnic diversity using Crunchbase data

Abstract: This pilot explores gender and ethnic diversity in the digital economy using a Crunchbase dataset of 98,231 unique companies and 145,736 unique personnel (e.g. employees, executives, board members) across 40 countries in Europe. We explore gender and ethnic diversity across job categories, degree type, and job roles. Analyses are performed at country level, in addition to a small selection of indicators computed at city-level in the UK. Women are underrepresented in the digital economy across all countries analysed. In the four countries with over 8,000 personnel in the dataset (i.e., United Kingdom, Germany, France, Spain), women are particularly underrepresented in 'software', whereas they are more likely than men to work in 'health care' and 'commerce and shopping' firms. We use a predictive model on first and last names of personnel to infer ethnicity. We then adapt the Simpson diversity index to obtain a single-value indicator for ethnic diversity at country level. Digital industries in several countries (e.g. Italy, Russia) are characterised by a low level of ethnic diversity, whereas a number of countries exhibit a broader mix of ethnicities in the digital industries (e.g. Finland, Estonia and Switzerland). We adapt the Lieberson diversity index to develop a single-value indicator combining gender and ethnic diversity at country and city level. A key challenge identified, which can be further explored should this pilot scaled up, is how to contextualise and benchmark these findings against other data sources. This has the potential to be particularly challenging for analyses of ethnic diversity, data for which are inconsistent across Europe.

1 Introduction

1.1 Background/context

'Inclusion' has risen to the forefront of numerous political, social and economic agendas globally in recent years. Growing concerns over the destabilising effect of inequalities within and between countries, as well as political and socioeconomic rifts that have grown deeper amidst unparalleled technological progress, have underpinned this agenda.

Despite widespread variability in the definition and measurement of this concept, its surging popularity suggests a societal desire to rethink the systems that have led to social disenfranchisement, political unrest and environmental degradation. In order to usher in a new era of truly inclusive innovation, new empirical analyses are needed that will allow decision-makers to develop, implement and evaluate policies in this domain.

Inclusive innovation: an ever-evolving concept

'Inclusive innovation' is a concept that has evolved from a series of interrelated paradigms. Initially emerging out of bottom-of-the-pyramid and poverty alleviation frameworks, the concept was later broadened to consider socially excluded groups more generally (Chataway, Hanlin, and Kaplinsky 2014). Inclusive innovation has also been defined as "the means by which new goods and services are developed for and by marginal groups (the poor, women, the disabled, minorities, etc.)" (Heeks, Amalia, and Shah 2013). The concept of inclusion is also invoked in the responsible research and innovation

paradigm, where it is defined as the involvement of broader stakeholder groups within science and policy as a means of legitimation (Stilgoe, Owen, and Macnaghten 2013).

Diversity as a core tenet of inclusive innovation in firms

The concept of diversity is deeply intertwined with that of inclusion. However, a recurring tension between diversity as a means of spurring micro- or macro-level economic performance (i.e. a 'business case'), or as a matter of social justice, is apparent in streams of management and innovation literature (Tatli and Özbilgin 2012; Bozeman and Sarewitz 2011).

The 'business case' argument is borne out of an attempt to develop demand-side cases for equality, appealing to the self-interest of individual employers.[1] In many countries, firms have responded by aggressively adopting diversity and inclusion strategies (Mayer, Warr, and Zhao 2018).

Although firm-level studies exploring links between diversity and innovation performance are relatively rare, evidence does suggest a positive association. For instance, a 2018 study of United States (US) firms found that those with corporate policies and pro-diversity cultures (as measured through a wide range of indicators such as having a woman CEO) enhance future innovative efficiency (Mayer, Warr, and Zhao 2018). The effects were more pronounced during economic downturns, and also in firms that had stronger governance, higher cash flow, greater growth options, are (already) more innovative, and place higher value on intangibles and human capital. The authors of this study employed a Granger causality regression to changes in new product announcements on lagged change in diversity policies, finding results consistent with causality (Mayer, Warr, and Zhao 2018). They also test the idea that firms located in California - which is a highly innovative and socially progressive state - may be driving the effects, concluding that findings are valid both within and outside of California.

Similarly, a 2011 study analysed a linked employee-employer dataset of over 1,600 firms in Denmark, finding a positive association between diversity in education and gender on the likelihood of introducing an innovation (Østergaard, Timmermans, and Kristinsson 2011). A positive association was also identified between an open culture towards diversity and firm-level innovative performance. However, age diversity negatively impacted firm-level innovation and no significant effect of ethnicity was noted on the firm's likelihood to innovate. The authors suggest that the lack of effect from ethnic diversity "might be explained by the high share of Danes among the employees", or that "a higher share of foreigners might take routine type work with low entry barriers in low innovative industries or in highly innovative firms that are sourcing very specialised employees regardless of ethnicity" (Østergaard, Timmermans, and Kristinsson 2011).

Although representing a different facet of the innovation ecosystem (i.e. entrepreneurs rather than firms), findings from studies in the US and United Kingdom (UK) have found that minority ethnic inventors play important roles in innovation (Nathan 2015).

The 'business case' argument for inclusion - particularly of women - has been extensively critiqued in the feminist economic literature (Thomson 2009). These critiques reflect, in part, a challenge to the dominant neoclassical narrative of productivity that determines how gross domestic product is measured, specifically in regard to the exclusion of the unpaid work performed largely by women in

[1] Most developed economies frame labour market outcomes for men and women as either a product of women's free choice (as rational economic actors), or more commonly as a lack of women's human capital formation (Thomson 2009). In response, most policy responses are supply side, aimed at the women themselves (e.g. availability of flexible working, market-based childcare, human capital investment, etc.) combined with anti-discriminatory legislation.

the household. In this system of measurement, women joining the labour force who take low paid, low quality jobs are contributing to productivity. Still, Thomson (2009), on examining the feminist economic perspectives on the business case for gender equality in the UK labour market, highlights that "it or may not be good for business, depending on the industry in question or prevailing economic conditions, but [gender equality] is always and for every industry and occupation, moral and just" (Thomson 2009).

Inclusive innovation policy

The OECD broadly defines inclusive innovation policies as those that "aim to remove barriers to the participation of individuals, social groups, firms, sectors and regions that are underrepresented in innovation activities in order to ensure that all segments of society have the capacities and opportunities to successfully participate in and benefit from innovation" (Planes-Satorra and Paunov 2017).

Delving into the social aspect of this definition, it becomes apparent that contextual differences are already codified in the way in which 'diversity' is interpreted, and policies enacted, across countries. For instance, in Scandinavian countries, the term 'diversity' typically refers to ethnic minorities (Tatli and Özbilgin 2012). Comparing the innovation policies of Germany and Israel, Zehavi and Breznitz (2017) find that in Germany, much effort has gone into supporting policies to advance women's employment or academic and business career but little effort on integrating disadvantaged minorities, while in Israel the reverse was true (Zehavi and Breznitz 2017). Cross-country differences are perhaps most clearly illustrated by Table 3, which shows the variable combinations of social groups targeted within innovation policies of 10 countries.

Table 3: Social group participation targets in innovation policies

Country	Women	Ethnic minorities/ immigrants	Low-income/ economically marginalised	People with disabilities
Brazil			●	
Canada	●	●	●	●
Chile				
France				
Germany	●	●		
Israel	●	●		
Norway	●			
South Africa	●	●	●	●
Sweden	●	●	●	
United Kingdom	●	●	●	●

Source: Reproduced from *How inclusive is innovation policy?* (Stanley, Glennie, and Gabriel 2018)

1.1.1 Opportunity

As a growing chorus of actors in the private and public sectors call for more inclusive innovation, empirical evidence with which to develop, implement and evaluate policies and programmes must keep pace (Schillo and Robinson 2017; Stanley, Glennie, and Gabriel 2018). The ground is therefore fertile for new indicator development in this domain. For instance, intersectionality[2] is a concept that has yet to be deeply explored in the innovation literature.

This pilot takes advantage of a novel, firm-level dataset, as well as a suite of new methods and tools, to develop indicators on gender and ethnic diversity in the digital economy in Europe. It surpasses traditional diversity analyses by also adapting from ecology two types of diversity indices that facilitate cross-context comparisons.

1.1.2 Flexibility of application domain

The analyses performed in this pilot are application domain-agnostic, and can therefore be broadly applied. The greater challenge of applying this method elsewhere is that it requires individual- and firm-level data.

1.1.3 Assessment of 'periphery to core of R&I policy' capacity of pilot

The use of Crunchbase in this pilot increases the likelihood of the indicators moving from the periphery to the core of research and innovation (R&I) policy, given its high level of geographic coverage and growing acceptance amongst academics and other stakeholder groups (Dalle, den Besten, and Menon 2017; Nathan, Kemeny, and Almeer 2017; Tarasconi and Menon 2017). However, in order to move to the core of R&I policy, additional quality checks and validation exercises would need to be undertaken to ensure the robustness of findings derived (see Section 4.2 for further details).

1.1.4 Stakeholder engagement summary

This pilot has benefitted from a wide variety of stakeholder engagements both within and outside the context of EURITO. These include:

- Ongoing engagement with the Inclusive Innovation team at Nesta. At the time of writing, an event was being planned for April 2019 to explore opportunities for new data and analyses for innovation, where the work from this EURITO pilot will figure prominently.
- The EURITO Knowledge Stakeholder Workshop provided an opportunity to discuss various pilot options within the 'inclusive' stream. For instance, a representative from the European Commission's Directorate-General for Research and Innovation Unit B7 (Science with and for Society) was particularly interested in championing pilots on gender equality in R&I.
- Nesta has been engaging with members of the UK Innovation Caucus - an initiative funded by Innovate UK and the Economic and Social Research Council. This group is currently developing a line of work on black, asian and minority ethnic groups in UK innovation.
- As part of a parallel strand of work to develop inclusive innovation indicators in Scotland, the Nesta team has been engaging with various stakeholders from the Scottish Government, Scottish Enterprise, and the Scottish higher education system. A workshop in fall 2018 provided an opportunity to gather stakeholder feedback on the topic of inclusive innovation indicators.

[2] Intersectionality is a concept borne out of feminist legal theory and civil rights, and describes the way in which identities are compounded to produce effects that are not simply the sum of their parts (Else-Quest and Hyde 2016). For example, the experience of being a black lesbian is not simply the addition of black and gay cultures.

1.2 Relevance to RITO criteria

1.2.1 Relevant

As described above, the topic of inclusive innovation has increasingly become one of interest to a wide variety of stakeholders. However, empirical evidence to support policy development and evaluation in this domain remains relatively scant. Additionally, the concept of intersectionality remains underexplored in the innovation literature.

1.2.2 Inclusive

This central focus of this pilot is inclusion.

1.2.3 Timely

Crunchbase is frequently updated, providing the possibility of real-time (or near real-time) analyses.

1.2.4 Trusted

Crunchbase is increasingly being used to analyse innovation, and is increasingly appearing in analyses published in peer-reviewed publications and grey literature from highly respected sources (Dalle, den Besten, and Menon 2017; Nathan, Kemeny, and Almeer 2017; Tarasconi and Menon 2017). Still, in order for the analyses to be fully trusted at a wider scale, additional validation and quality checks should be undertaken (see Section 4.2).

1.2.5 Open

In line with EURITO's mandate to produce open indicators, notebooks with all of the code and outputs produced for this pilot are available through Nesta's inclusive innovation GitHub repository. The data used in the pilot can be licensed from Crunchbase.

1.3 Research/policy questions

This pilot addresses a range of research questions:

- How ethnically and gender-diverse are digital firms in Europe?
- How do ethnic and gender diversity vary across company categories, job types, and degree types?
- Does the ethnic composition of men and women the digital economy differ (this is an indicator of intersectionality, as explained above).

2 Methodology

Several terms used in this paper are defined below for clarity:

- 'Gender diversity' is conceptualised as the relative proportion of women and men in a given geographic or sectoral boundary.
- 'Ethnic diversity' is conceptualised as the relative representation of different ethnic groups within a given geographic or sectoral boundary.
- 'Personnel' is the term used to refer to all individuals in the Crunchbase dataset represented by a unique person ID. These include employees, executives, board members, board observers, or advisors. A large proportion of the personnel in Crunchbase are executives.

2.1 Data sources

Crunchbase is a 'frontier' dataset that is increasingly being used to explore the digital economy (Nathan, Kemeny, and Almeer 2017). Founded in 2007, Crunchbase initially tracked firms that appeared on the TechCrunch industry news site. It has since evolved into a wide-reaching, firm-level crowdsourced dataset with rich information on technology-oriented firms, founders, employees, investors and investments (Nathan, Kemeny, and Almeer 2017). The dataset is regularly updated and has near-global coverage (containing firms in over 200 countries, although country-level coverage is uneven), making it a valuable source of insight for large-scale analyses of the digital economy.

Data in Crunchbase are collected through a combination of crowdsourcing and curation. This approach has the potential to introduce biases into the broader dataset, for example through incorrect or intentionally misleading personal data entry.

2.2 Methods

Tools: All analyses were carried out using Python version 3.6. The following packages and libraries were used (also shown in the requirements.txt file):

- Ethnicolr: An open source, Python-based predictive model of race/ethnicity based on first and last name. The model was trained on Florida voting registration data and Wikipedia data. The relationship between the sequence of characters in a person's name and their race/ethnicity is modeled using Long Short Term Memory Networks. For full name estimation on the Florida Voter Registration data, the model's out of sample precision is 0.83 and its recall is 0.84. The model is explained in more detail in the 2018 paper by Sood and Laohaprapanon (Sood and Laohaprapanon 2018), and its potential limitations in this study are described in Section 4.2.
- NumPy: A core package for scientific computing in Python.
- Pandas: Python data analysis library.
- Scikit-bio: A Python library for bioinformatics, containing a wide range of diversity metrics.
- Matplotlib: 2D Python plotting library used for basic visualisations.

Data collection: The data used in this pilot were collected via the Crunchbase API. Relevant tables were merged to obtain the final dataset used in the analyses. These tables are: Organisations, category groups, geographic data, degrees, jobs, and people. The final dataset contains the following variables: unique organisation ID, total funding (United States Dollars), company founding date, city, country, employee count, unique person ID, first name, last name, gender, race, degree type, degree ID, (educational) institution ID, primary role, job ID, job type, category group list. The final dataset contains 98,231 unique companies and 145,736 unique personnel across 40 countries - the EU-28 as well as other countries on the European continent.

Exploratory data analysis: Exploratory data analysis was performed to assess missing data. Personnel education data (degree type, degree ID, institution ID) is missing in over 50% of the observations, while total funding in USD is missing for 60% of companies. Employee count is missing for 15.9% of companies, gender (15.2% missing), full name (13.2% missing), followed by category group (9.4% missing), company founded date (7.5% missing) .

Indicator development and analysis: Analyses on the number of companies and personnel per country show - where needed - unique variables only (i.e. dropped duplicates).[3] The following set of indicators were developed for this pilot:

- Gender/ethnic diversity: The proportion of men/women and the breakdown of ethnic groups in the Crunchbase dataset (within a given city or country).
- Gender/ethnic diversity in roles: The proportion of men/women and ethnic groups represented in the Crunchbase dataset, by role (i.e. job type, such as 'executive'), by city and country.
- Gender/ethnic diversity in degrees: The proportion of men/women and ethnic groups represented in the Crunchbase dataset, by degree type (i.e. undergraduate, postgraduate, MBA, PhD). Note that one person can hold multiple degrees.
- Studied in same country: The proportion of personnel who attained a degree in the same country in which their organisation is located. Note that this indicator should be cautiously interpreted given a high level of missing data.
- Gender/ethnic diversity in company categories: The proportion of men/women and ethnic groups represented in the Crunchbase dataset, by company category (i.e. Crunchbase standard categories[4], such as software, finance, etc.).
- Intersectional indicator of participation in the digital economy: An indicator showing relative ethnic diversity disaggregated by gender.[5]
- Diversity indices:
    - $Simpson\ Index = 1 - \Sigma {p_i}^2$ where p_i is the proportion of the community represented by i. This index[6] is a measure of diversity that takes into account the number of entities present in a given location, as well as the relative number of each entity. The result is an index ranging from 0 to 1, with 1 representing maximum diversity and 0 representing no diversity. (McLaughlin et al. 2016). In this pilot, we use it to measure the gender and ethnic diversity on city/country-level.
    - $A_w = 1 - (\sum_{k=1}^{p} \frac{Y_k^2}{V})$ where Y_k is the proportion of the population falling in a given category within each of the variables, V is the number of variables and p is the total number of categories within all of the variables. When using Lieberson index (Sullivan 1973), the result ranges from 0 to 1, with 1 representing infinite diversity and 0 representing no diversity. In this pilot, we combined gender and ethnic diversity on city/country-level to measure it.

2.3 Documentation

In line with EURITO's mandate to produce open indicators, notebooks with all of the code and outputs produced for this pilot are available on Nesta's GitHub repository. The data used in the pilot can be licensed from Crunchbase.

[3] For the degree related indicators, duplicates of `person_id` were retained because a person can hold multiple degrees.

[4] Crunchbase categories are assigned by verified users of the platform. See Section 4.3 for further considerations on next steps for categorising company industries.

[5] Some have argued that qualitative methods are more appropriate for studying intersectionality, although the use of quantitative methods has been recognised in sociology, gender studies and family studies. There is substantial diversity across theorists and researchers in what is considered to be an intersectional approach or intersectional analysis (Else-Quest and Hyde 2016).

[6] Simpson's Diversity Index originated in ecology as a means to quantify the biodiversity of a habitat. In ecological terms, the inputs to the equation are the number of species present (termed 'richness'), and the relative abundance of different species present (termed 'evenness').

3 Results

3.1 Outputs

Figure 1 shows the number of unique European companies by country in Crunchbase. The majority of companies are in the United Kingdom (n=31,591), followed by Germany (n=10,041), France (n=9,454), Spain (n=6,929) and the Netherlands (n=5,044).

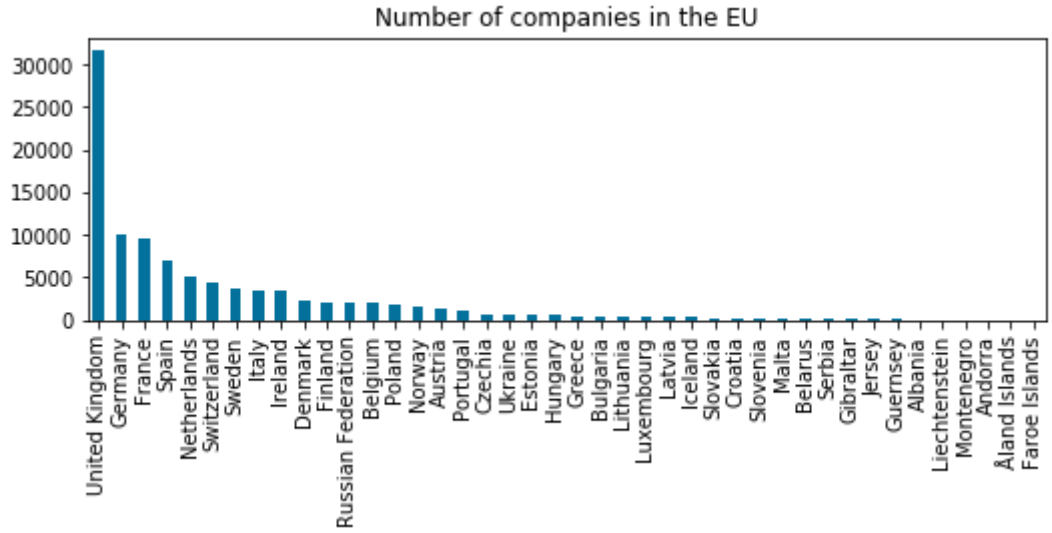


Fig. 1 Number of digital companies in Europe, by country

Figure 2 shows the number of personnel in European digital companies, demonstrating a very similar trend to Figure 1. The United Kingdom has the most personnel (n=50,708) listed in Crunchbase, followed by Germany (n=16,189), France (n=12,440), Spain (n=8,436), and the Netherlands (n=7,518).

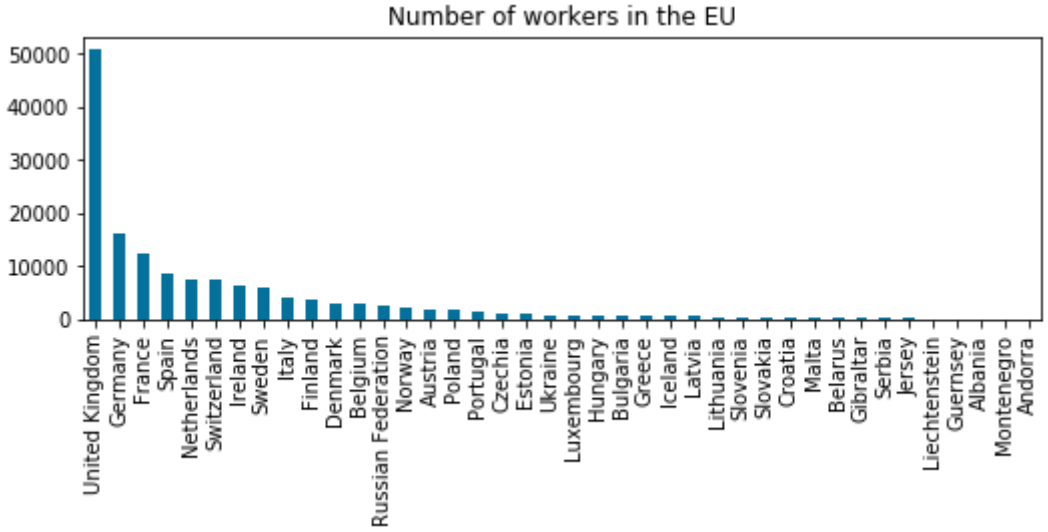


Figure 2: Number of personnel in digital companies in Europe, by country

Figure 3 shows the size of companies in Crunchbase across European countries, organised by the date of their founding, between 1990 and 2018. Note that the starting date of 1990 is selected for visual clarity in the figure. Micro-firms (10 or fewer employees) comprise the largest proportion of firms, followed by small firms (between 10 and 50 employees). The annual rate of company establishment peaked in 2013, nearing 4,000 new companies in that year.

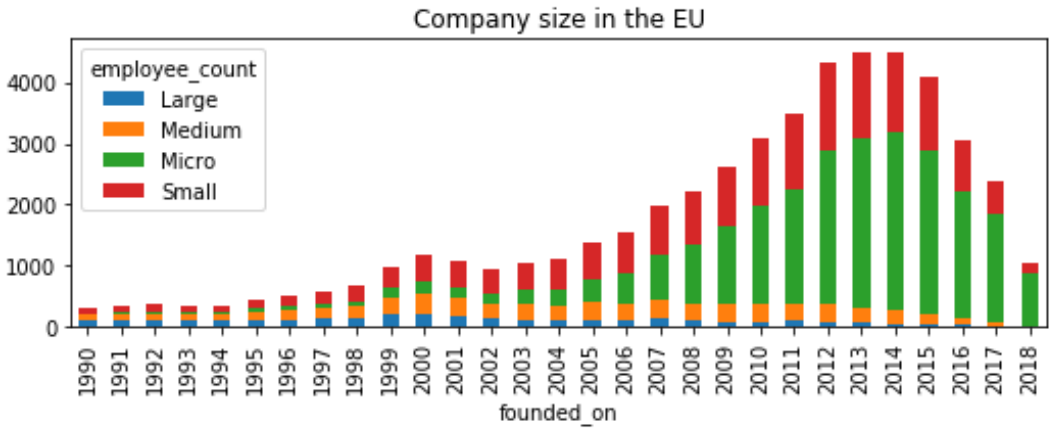


Figure 3: Size of European companies in Crunchbase

Figure 4 shows the amount of funding in billions of US dollars[7] (through instruments such as seed, angel, and venture funding) European companies in the digital economy have received between 1990 and 2018. Note that the starting date of 1990 is selected for visual clarity in the figure. Annual funding reached its highest level in 2007, before dropping off in 2008-2009, and then rebounding for several years that followed. Note that funding in USD is missing in 60% of observations, so findings should be cautiously interpreted.

[7] Not adjusted for inflation

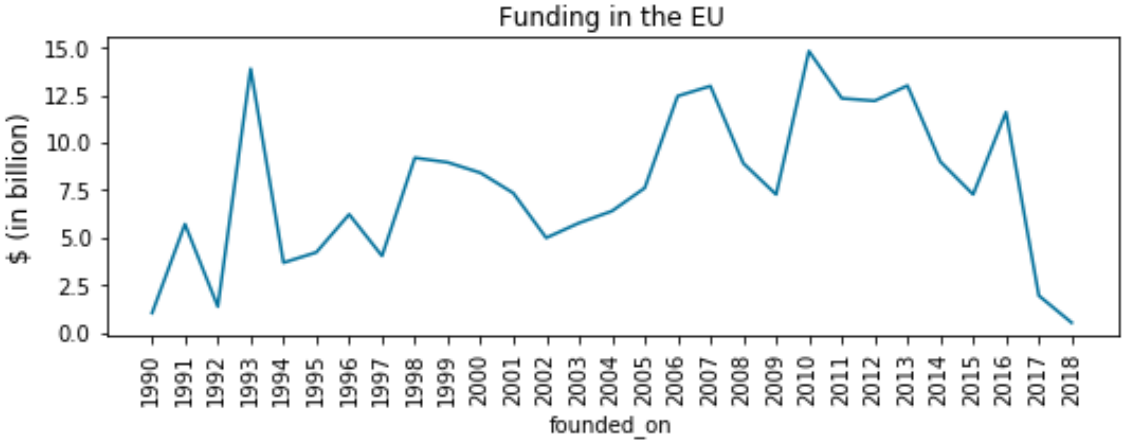


Figure 4: Funding of European companies in Crunchbase

3.2 Findings

3.2.1 Gender and ethnic diversity across countries

Across European countries in Crunchbase, personnel in the digital economy are predominantly men, with none of the countries achieving over 20% women personnel (Figure 5). The proportion of women personnel is highest in the United Kingdom, where women make up 16.9% of personnel, followed by Ireland (15.9%), Spain (15.8%) and Sweden (15.4%).

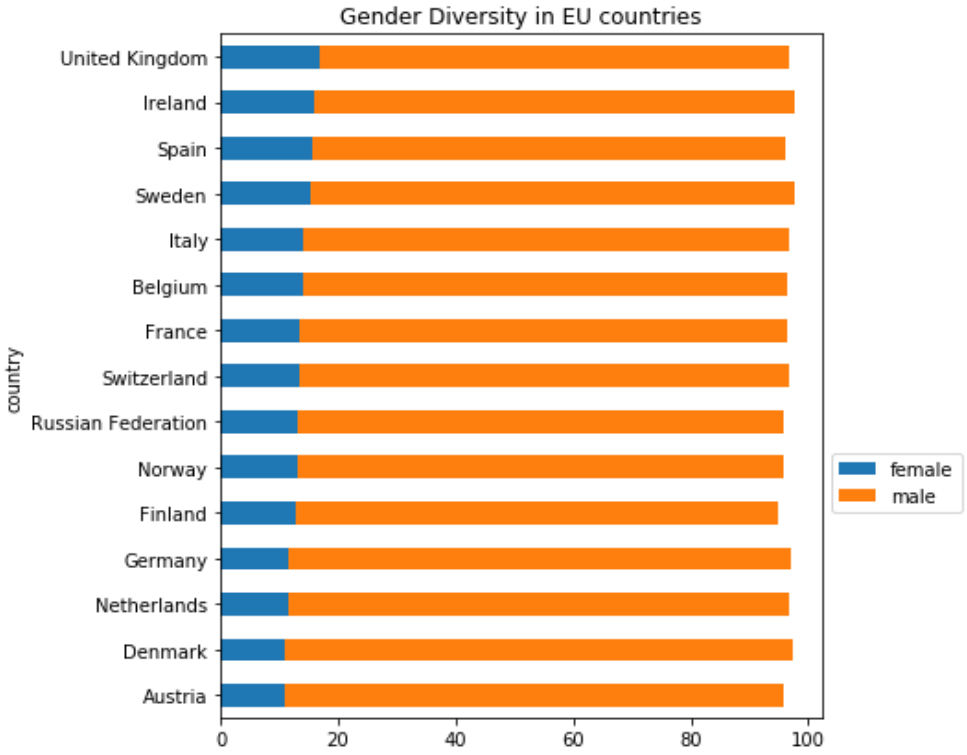


Figure 5: Gender diversity in European digital economy

Figure 6 shows the distribution of ethnic composition of personnel in Crunchbase by country. In several countries, a notable predominance of one ethnic group is apparent. For example, in Italy,

82.6% of personnel are classified as 'Italian' ethnicity and in Russia, 82.8% of the personnel are classified as 'Eastern European' ethnicity.

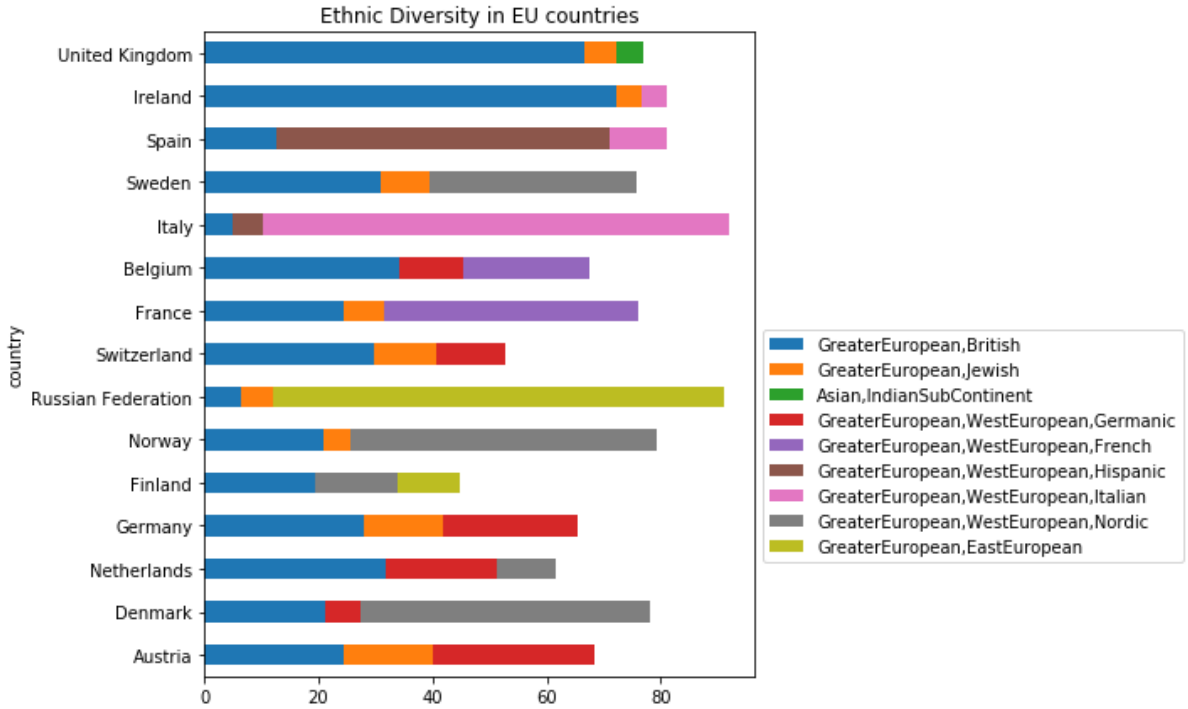


Figure 6: Ethnic diversity in European digital economy

In order to facilitate the interpretation of ethnic diversity across countries, we compute the Simpson Index. A score of 0 is indicative of a low level of ethnic diversity, whereas a score of 1 indicates complete ethnic diversity. The Simpson Index of ethnic diversity is shown in Figure 7. By this metric, Finland has the highest ethnic diversity (0.885), followed by Switzerland (0.842).[8] By contrast, Italy has the lowest Simpson index (0.371), followed by Portugal (0.403) and Russia (0.408).

Across Europe, data collection on race and ethnicity has not been standardised at regional or even national level in many cases (Farkas 2017), making a straightforward comparison of the findings difficult. This challenge is discussed further in the Limitations section of the paper.

[8] It may be argued that not all 'types' of diversity should be weighed equally (e.g. the diversity obtained by employing multiple 'Greater European' ethnicities is not the same as that of, for example, 'Indian Subcontinent' and 'Greater European'. Although this argument is compelling and deserves further consideration, it is beyond the scope of a pilot paper.

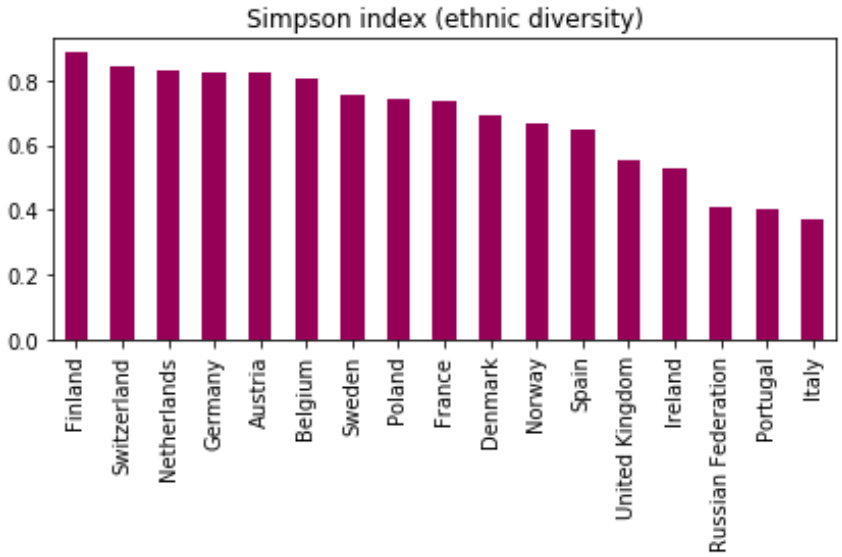


Figure 7: Simpson Index of ethnic diversity

Figure 8 shows the proportion of ethnic groups across genders in the four largest countries (by number of personnel) in the dataset. Men and women personnel in the UK exhibit approximately the same ethnic composition, while in Germany, France, and Spain, the most common ethnic group in each country (e.g. 'French' in France) is more predominant amongst men than women. In all countries except the UK, the proportion of 'British' women is higher than amongst men. This finding should be further unpacked and explored in future analyses.

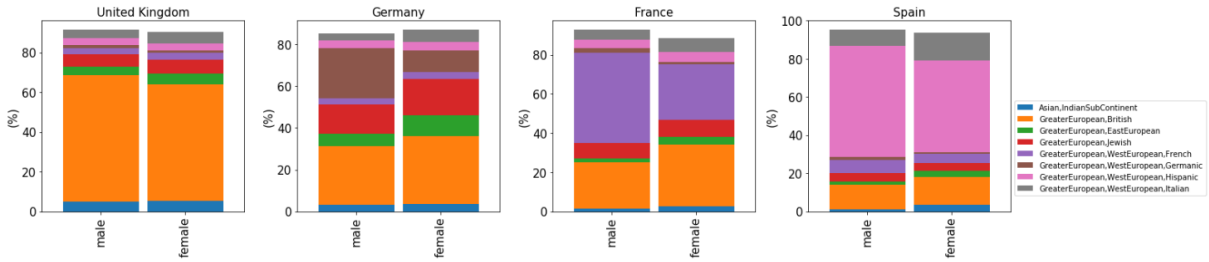


Figure 8: Intersectionality in European digital economy

3.2.2 Gender and ethnic diversity by company category

Figure 9 shows gender diversity across a selection of 'company categories' from Crunchbase. In a finding similar to the analysis at country level, the proportion of women personnel across all industries is far lower than men. Women are least represented in transportation (10%) and software (10.7%), and most represented in education (19.9%) and healthcare (18.1% women).

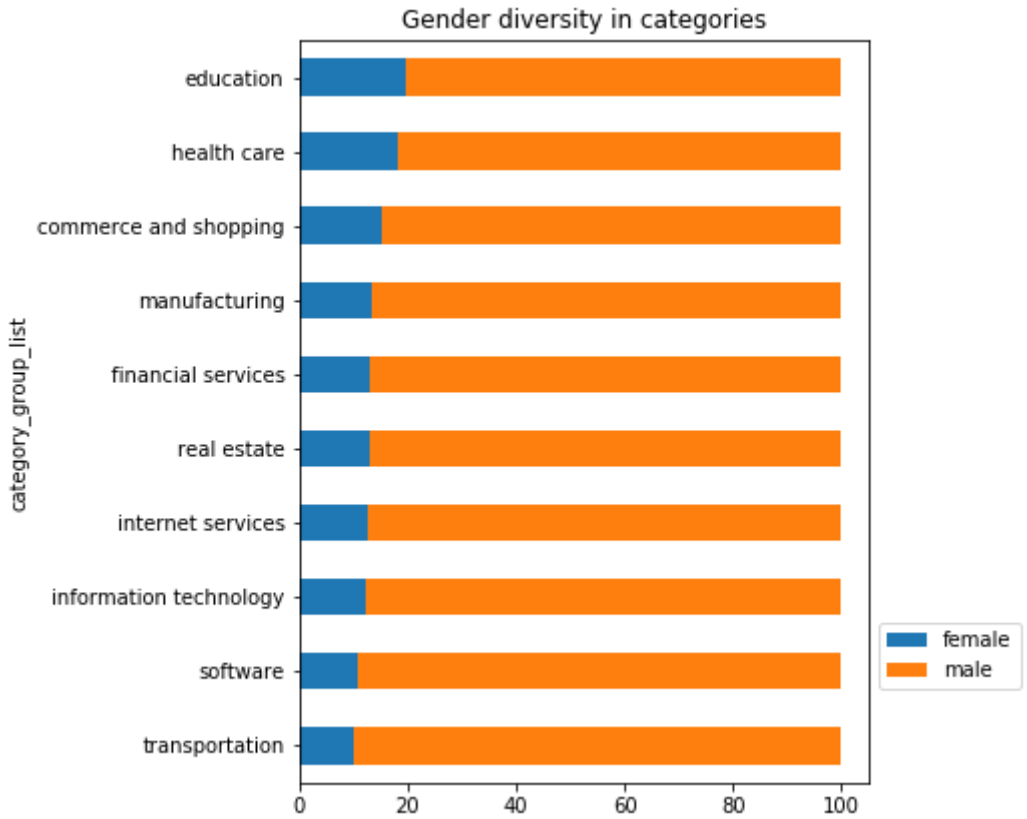


Figure 9: Gender diversity by company category

Figure 10 shows the gender diversity by country across the five most common Crunchbase categories. The figure shows the four countries with the most personnel in the dataset (i.e., UK, Germany, France, Spain, which all have over 8,000 personnel). In all countries, women are more likely than men to work in 'health care' and 'commerce and shopping', although the size of this difference varies across countries. By contrast, men are more likely to work in 'software' in all countries. Germany is the only country of the four in which women figure more prominently in 'information technology'. The lowest proportion of men and women in all countries is in 'manufacturing', with a slightly higher proportion of women in Germany and France.

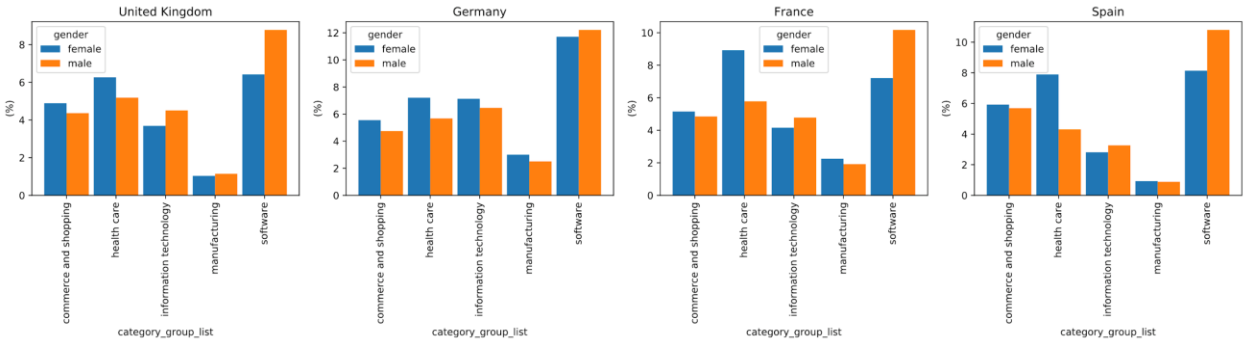


Figure 10: Gender diversity by country and company category

Figure 11 shows the ethnic diversity across the same categories as Figure 9. There are some minor variations across categories, however the ethnic distribution across categories is relatively consistent.

'British' ethnicity comprises the largest share (close to 40%) across all of the categories. This may be due simply to the fact that the UK has the highest number of companies and employees in the dataset.

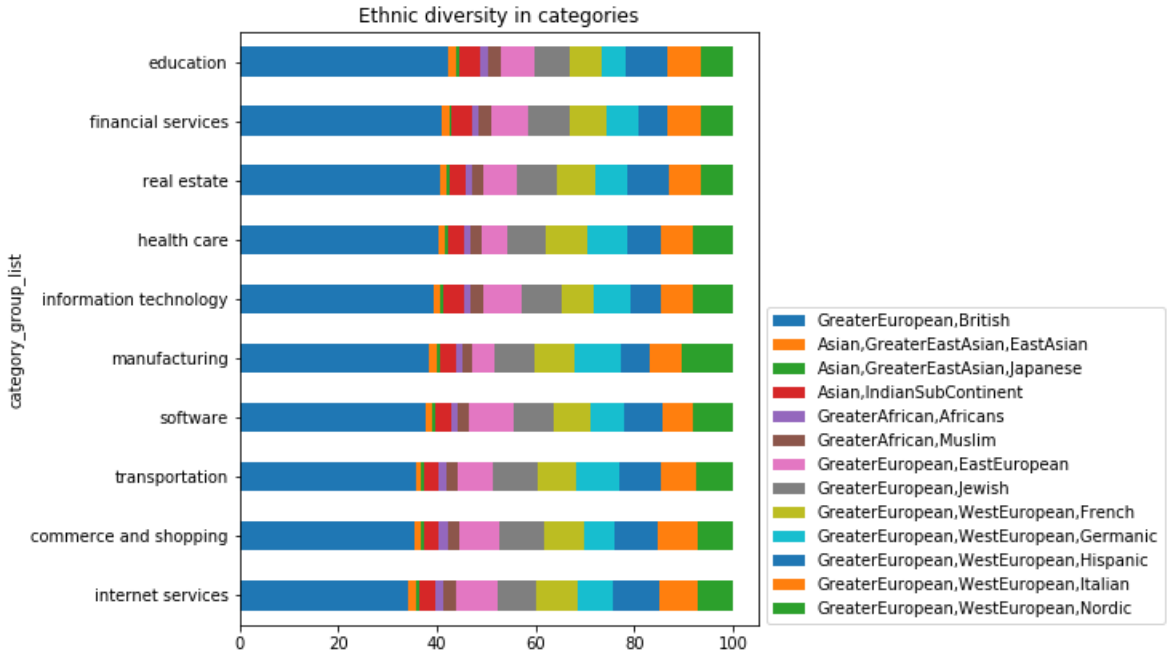


Figure 11: Ethnic diversity by company category

3.2.3 Education of personnel in Crunchbase

Figure 12 shows the gender diversity in degree types across the four top countries (by number of personnel) in Crunchbase. Across all countries, a higher proportion of women have postgraduate degrees than men. However, men are more likely to hold an MBA across all countries, and men are also more likely to hold a PhD (with the exception of Spain).

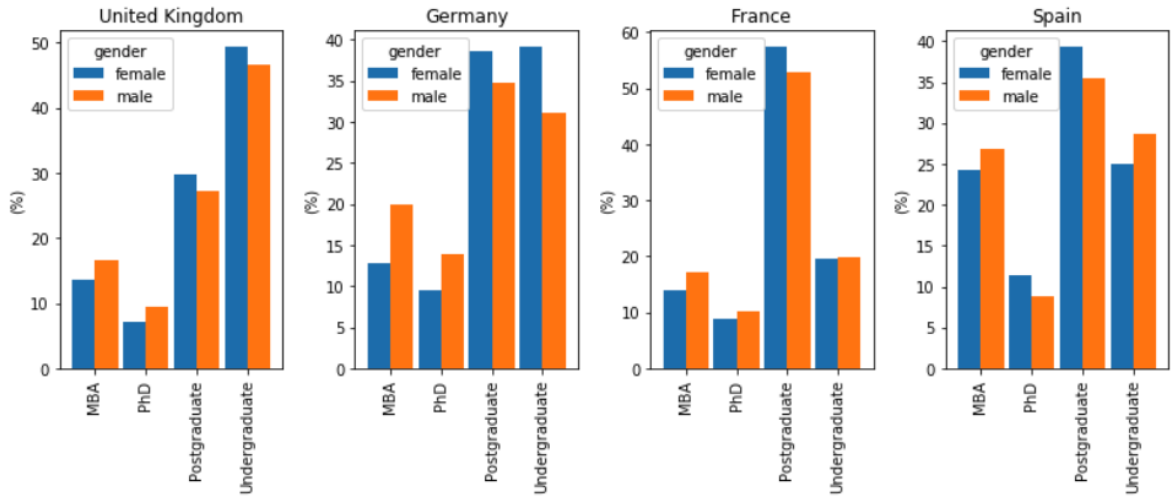


Figure 12: Degree type by gender and country

Figure 13 shows ethnic diversity by degree type for the three most common ethnic groups by country, in the three countries with the largest number of personnel (UK, Germany, France) in the dataset. In the UK, an undergraduate degree is the most common amongst all ethnic groups, however personnel of 'British' ethnicity are more likely than those of 'Jewish' or 'Indian' ethnicity to hold one. However,

this trend reverses for postgraduate degrees, which are most prevalent amongst personnel of 'Indian' and 'Jewish' ethnicities. In Germany, a larger degree of variability exists across ethnic groups, with 25.6% of personnel of 'Germanic' ethnicity holding PhDs, relative to 10.9% of 'British' ethnicity and 11.4% of Jewish' ethnicity. In France, postgraduate degrees are far more common than MBAs or PhDs. Personnel of 'French' and 'Jewish' ethnicities are more likely to hold a postgraduate degree (64% and 63.8%, respectively) than those of 'British' ethnicity (43%), who are themselves more likely to hold an MBA or undergraduate degree.

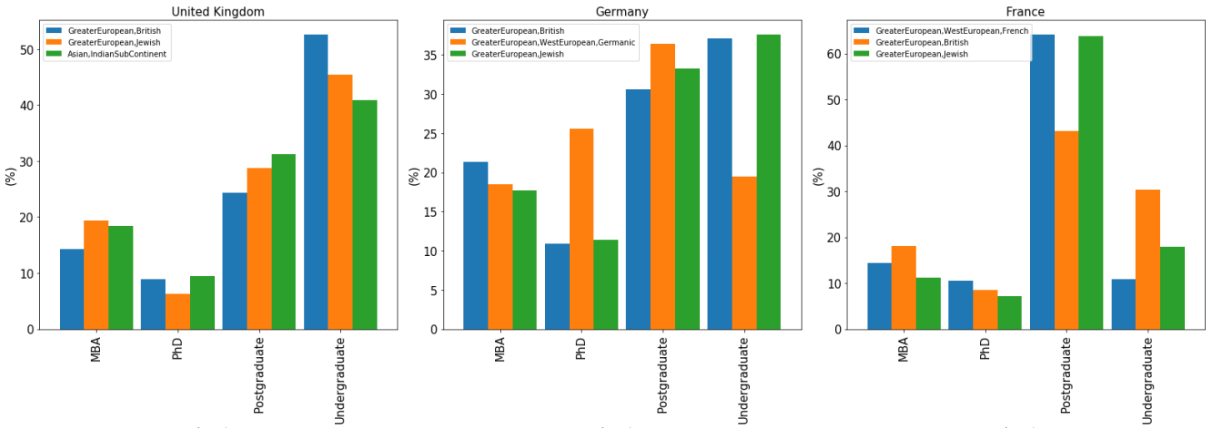


Figure 13: Ethnic diversity by degree type

Figure 14 shows the proportion of personnel who studied in the same country as where their company is located. Spain (15.8%) has the highest proportion of locally-educated personnel, followed by Norway (15.7%), France (15.4%), Denmark (14.8%), and Sweden (14.5%), while Luxembourg, Lithuania and Ireland have comparatively lower levels of locally-educated personnel (under 5%). Several countries which have very few personnel in the dataset have no locally-educated personnel.

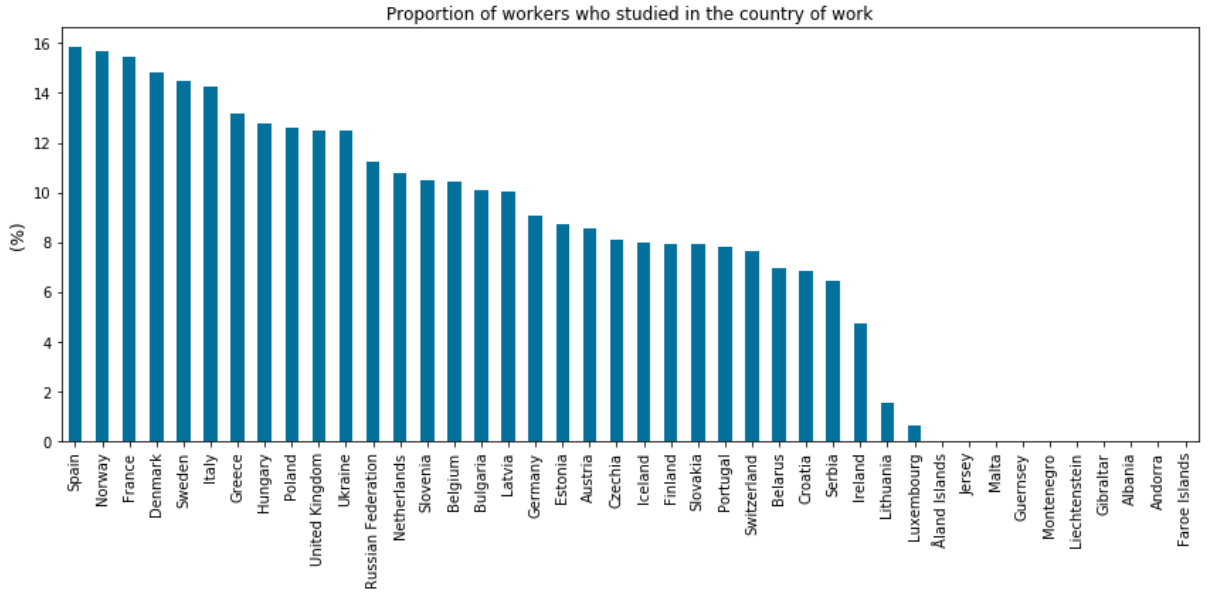


Figure 14: Proportion of personnel who studied in the country

3.2.4 Gender and ethnic diversity in role type

Figure 15 shows gender diversity by role type across the four largest countries (by number of personnel) in the dataset. Across the countries, men are more likely than women to be in executive and board member roles, while a higher proportion of women are classified as employees.

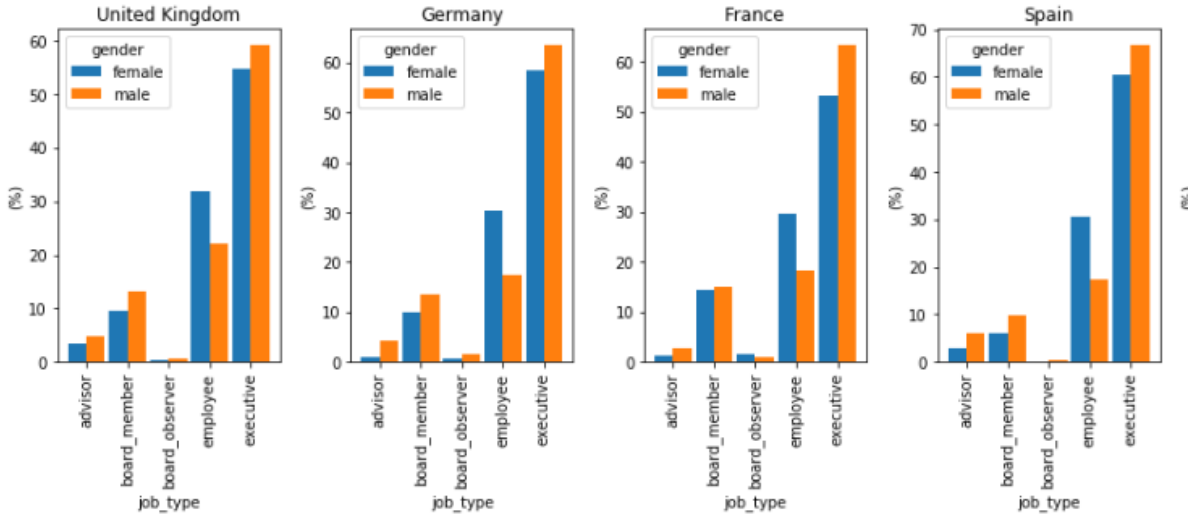


Figure 15: Gender diversity by role type

Figure 16 shows ethnic diversity by job type for the three most common ethnic groups by country, in the three countries with the largest number of personnel in the dataset (UK, Germany, France). Across the three countries, the distribution of personnel across categories, as well as the ethnic diversity within each category, is relatively consistent.

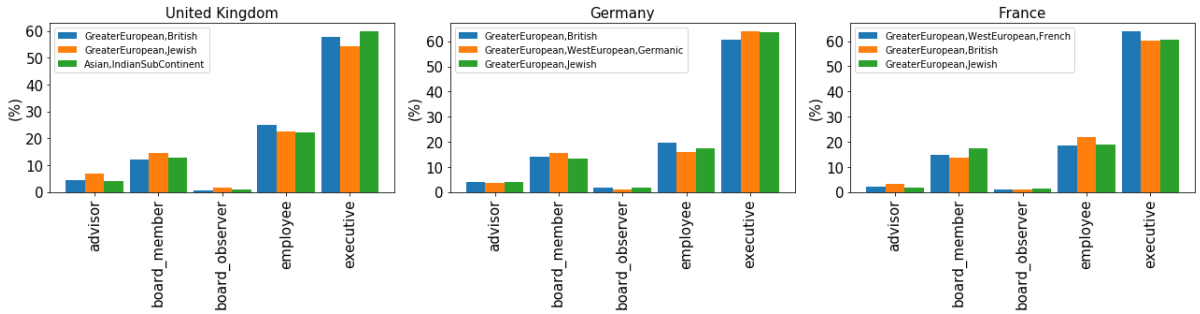


Figure 16: Ethnic diversity by role type

Note: Colour scheme of ethnicities differs across countries

3.2.5 Lieberson index of gender and ethnic diversity

The Lieberson index combines the gender and ethnic diversity of countries into a single value. The index is presented on a scale of 0 to 1, where a score of 0 represents a low level of diversity and a score of 1 represents a high level of diversity.

Figure 17 shows the Lieberson Index across countries in Europe. Finland performs the best on the index (0.55), followed by Switzerland (0.54), and the Netherlands (0.51). Italy (0.30), Portugal (0.32) and the Russian Federation (0.32) have the weakest performance in the index. As with Italy and the Russian Federation, Portugal's employee ethnicity is highly dominated by one group (in this case, Hispanic).

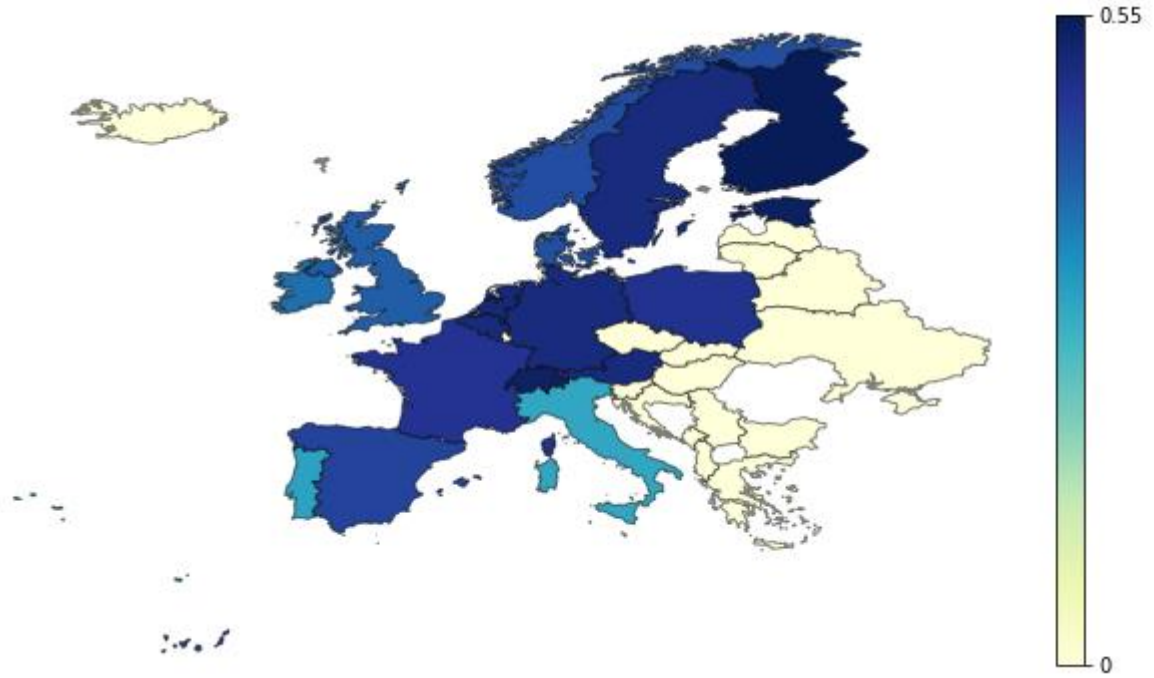


Figure 17: Lieberson Index across European Countries

3.2.6 Country case study: the United Kingdom

Figure 18 shows gender diversity by degree type in the four largest UK cities (by personnel number) in the dataset. In London and Edinburgh, over 40% of men and women hold undergraduate degrees. The proportion of personnel with a PhD is far higher in Cambridge (28% of women, 31% of men) than in Manchester, Edinburgh or London. In Cambridge, Manchester, and Edinburgh, women are more likely than men to hold an MBA.

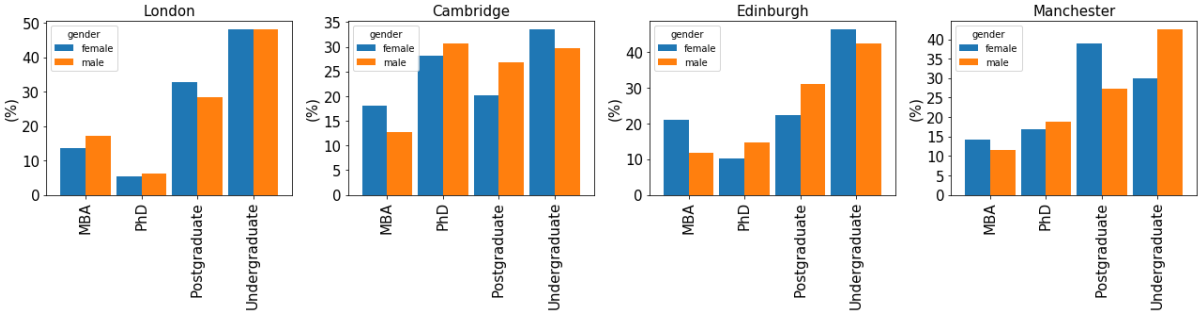


Figure 18: Gender diversity in degrees by city

Figure 19 shows ethnic diversity by degree type in the three largest UK cities (by personnel number) in the dataset. In London, personnel of 'British' ethnicity are most likely to hold an undergraduate degree, however those of 'Jewish' and 'Indian' ethnicity are slightly more likely to hold a postgraduate degree or PhD. In Cambridge, there are notable variations in ethnic groups across degree types. Personnel of 'Jewish' ethnicity are more likely to hold an undergraduate degree or MBA, however they are less likely than both 'Indian' and 'British' personnel to hold a PhD or postgraduate degree. In Edinburgh, a larger proportion of 'Hispanic' personnel hold postgraduate degrees than those of 'Jewish' or 'British' ethnicity.

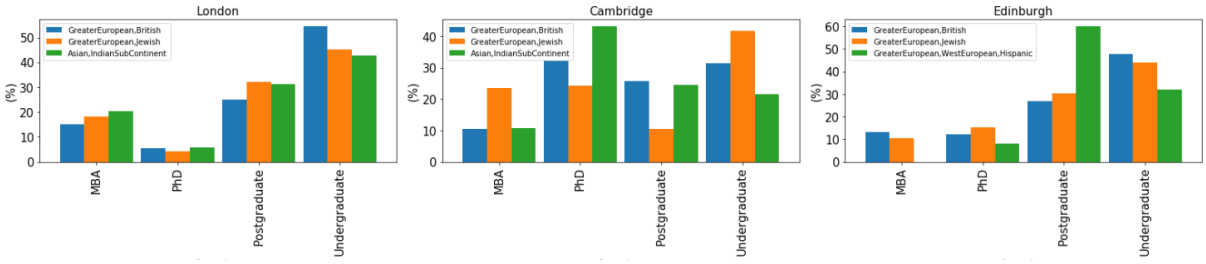


Figure 19: Ethnic diversity by city

Figure 20 shows the Lieberson Index across UK cities. London has the highest Liberson index (0.46), followed by Cambridge (0.39) and Reading (0.35). Bristol has the lowest Lieberson index (0.28). London's lead position in diversity is consistent with the 2011 census results, which revealed that it is the most ethnically diverse region (in England and Wales), with 40.2% of residents identifying as Asian, Black, Mixed or Other ethnic group (Office for National Statistics 2018).

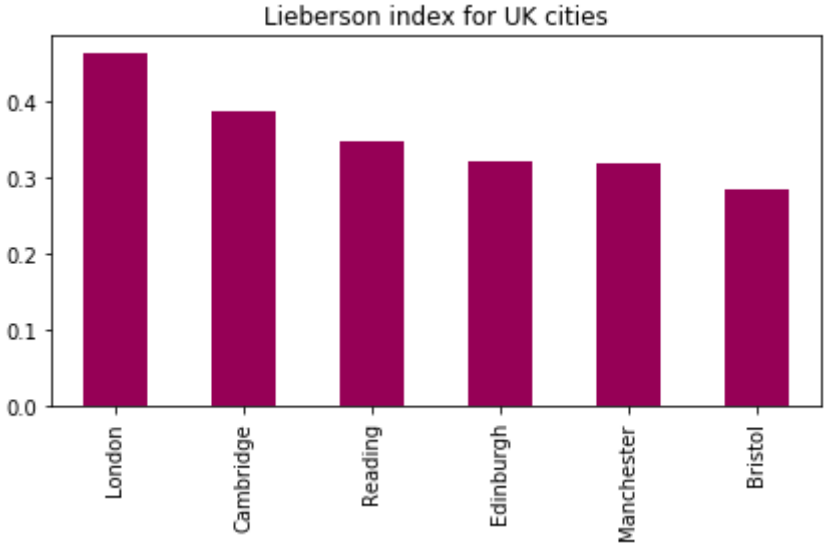


Figure 20: City-level Lieberson Index (United Kingdom)

4 Discussion and Conclusions

The analysis performed for this pilot reveals that across European countries, digital industries are primarily male-dominated. The proportion of women personnel is highest in the United Kingdom, where women make up 16.9% of personnel, followed by Ireland (15.9%), Spain (15.8%) and Sweden (15.4%). Across industry categories, women are least represented in software (10.7% women), and most represented in education (19.9%) and healthcare (18.1% women). Across countries, a higher proportion of women have postgraduate degrees, however men are more likely to hold a PhD or MBA than women. Board membership and executive positions across countries are also primarily male-dominated, whereas women make up a larger proportion of 'employees'.

Within several European countries (e.g. Italy), personnel in the digital economy are characterised by one dominant ethnic group, whereas in a selection of other countries (e.g. Finland and Switzerland), personnel in the digital economy are characterised by a mix of ethnic groups.

The Lieberson Index, which combines gender and ethnic diversity into a single value, largely mirrors the ethnic diversity of a given country. By this metric, Finland, Switzerland and Belgium have the most gender and ethnically-diverse digital economy, whereas Italy, Portugal and Russia have the least diverse. A lack of comparable race/ethnic data across European countries makes the contextualisation and benchmarking of these findings challenging (Farkas 2017), and is an area that requires further research.

4.1 Validation and ongoing stakeholder engagement

The results from this pilot will be validated with stakeholders in the months that follow the release of this paper. In particular, we will seek comments from Nesta's Inclusive Innovation team, as well as from a wider set of external stakeholders. The results of the pilot will also be showcased at an event titled 'new data for inclusive innovation' in April 2019. A parallel stream of work to develop new indicators of inclusive innovation within a Scottish context will also provide valuable opportunities to engage with stakeholders in this domain.

4.2 Limitations

Several limitations should be taken into account when reviewing this pilot research note. The first is that it is not yet clear to which extent Crunchbase captures the entirety of the digital economy in a given country. That is, do the wide variations in number of companies and personnel reflect the 'ground truth', or is this indicative of a bias within the dataset? In order to fully assess this potential limitation, findings should be validated against traditional indicators of digital economy activity activity across countries (e.g. using Eurostat data). A related validation step would be to explore and assess the data collection protocol for Crunchbase (i.e. what criteria are used to determine whether a company should be added to the database?). This would greatly improve the interpretability of the indicators.

An additional limitation is that the precision and recall of the ethnicity classifier has not yet been validated. This is particularly important given that the data were trained on a number of non-European datasets. One approach to validation would be to benchmark the performance of the Ethniclor classifier relative to a more robust tool such as the NamSor ethnicity classifier (which was trained on a dataset of names from the European Commission staff directory).

A further challenge regarding the ethnicity indicators was briefly touched upon in the Findings section (i.e. a lack of consistent, comparable ethnicity data across European countries against which the findings of this pilot can be benchmarked). In an in-depth assessment of ethnicity data across the European Union, Farkas (2017) finds that countries collect data on language, place of birth (of parents), or migration background. In large-scale, pan-European data collection exercises such as the Labour Force Survey and the European Statistics of Income and Living Conditions survey, proxy categories of citizenship and place of birth are collected (Farkas 2017).

4.3 Considerations for scaling up

4.3.1 Complementarities with other pilots

This pilot lays the foundation for subsequent analyses of inclusiveness across the R&I landscape, particularly in the digital economy. One option for scaling up would be to continue working with the Crunchbase data, for example to develop a machine learning model to impute missing values in company funding, or to make use of the 'event and conference attendance' data (not explored in this

paper) to examine diversity in attendance. It would also be feasible to analyse a subset of Crunchbase data for companies working on AI, making a clear link with the EURITO pilot on this topic.

Linking the Crunchbase data with another firm-level data source (e.g. company website data, patents) would also open another rich avenue toward further analyses (although the feasibility of this approach would need to be further explored, as it is not a trivial matter).

An additional step could be to cluster the categories or train a machine learning model on labelled website text data, using this to predict the market sectors of the Crunchbase companies.

An important component of scaling would be to attempt to better contextualise and benchmark the findings (particularly measures of ethnic diversity) against other country- or city-specific measures. While internationally-comparable data for some aspects of the analysis may permit such benchmarking (e.g. comparing gender equality in higher education or digital economy employment figured against Eurostat), this is anticipated to be much more difficult for other aspects of the analysis (e.g. ethnic diversity).

4.3.2 Tools and data sources

The tools and data sources used do not constrain the possibility of scaling up the analysis. Considerations for scaling up could, however, include linking Crunchbase with other data sources (e.g. web data), which would require additional planning and resources to implement.

4.3.3 Ethical considerations

While data science projects should always be subject to ethical scrutiny, this is essential for indicators that focus on individuals or collections of individuals, at the intersection of questions of vulnerability, discrimination and power structures. The question of ethics is therefore particularly salient in this pilot, and we must account for the fact that our capabilities to perform powerful analyses are evolving faster than our norms, rules and laws (Salganik 2017).

We must equally consider the risk of continuing under the status quo. If we do not pursue empirical studies of inclusive innovation, we risk implementing misguided policies that may not deliver the ultimate impact we seek. For example, increasing diversity within firms is a core tenet of inclusive innovation policy, however efforts may be misguided if we fail to understand the contextual variances and nuances of participation and outcomes. Additionally, by developing policies around historically marginalised groups (e.g. women) independent of the necessary data and indicators, we risk 'essentializing' (i.e. failing to recognise the heterogeneity *within* a given group) them, and ultimately failing to support those who would benefit most.

4.3.4 Conclusion

This pilot aims to move beyond the status quo in diversity measurement techniques by developing new indicators that capture - in a single value - ethnic and gender diversity in a given location. It also pushes beyond traditional approaches by developing an 'intersectional' indicator which allows for an assessment of ethnic diversity by gender. Most importantly, the pilot lays the foundation for further analysis of gender and ethnic diversity in the R&I landscape.

5 References

Dalle, Jean-Michel, Matthijs den Besten, and Carlo Menon. 2017. 'Using Crunchbase for Economic and Managerial Research'.

Farkas, Lilla. 2017. 'Data Collection in the Field of Ethnicity'. European Commission.

Heeks, Richard, Mirta Amalia, and Nishant Shah. 2013. 'Definition, Conceptualisation and Future Research Priorities'. *Development Informatics* Working Paper Series: 30.

Mayer, Roger C., Richard S. Warr, and Jing Zhao. 2018. 'Do Pro-Diversity Policies Improve Corporate Innovation?' *Financial Management*.

McLaughlin, Jacqueline E., Gerald W. McLaughlin, Josetta S. McLaughlin, and Carla Y. White. 2016. 'Using Simpson's Diversity Index to Examine Multidimensional Models of Diversity in Health Professions Education'. *International Journal of Medical Education* 7 (January): 1–5. https://doi.org/10.5116/ijme.565e.1112.

Nathan, Max, Tom Kemeny, and Bader Almeer. 2017. 'Using Crunchbase to Explore Innovative Ecosystems in the US and UK'.

Office for National Statistics. 2018. 'Regional Ethnic Diversity'. 2018. https://www.ethnicity-facts-figures.service.gov.uk/british-population/national-and-regional-populations/regional-ethnic-diversity/latest.

Planes-Satorra, Sandra, and Caroline Paunov. 2017. 'Inclusive Innovation Policies: Lessons from International Case Studies'. OECD Science, Technology and Industry Working Papers 2017/02. https://doi.org/10.1787/a09a3a5d-en.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Schillo, R Sandra, and Ryan M Robinson. 2017. 'Inclusive Innovation in Developed Countries: The Who, What, Why, and How'. *Technology Innovation Management Review* 7 (7): 13.

Sood, Gaurav, and Suriyan Laohaprapanon. 2018. 'Predicting Race and Ethnicity From the Sequence of Characters in a Name'. *ArXiv:1805.02109 [Stat]*, May. http://arxiv.org/abs/1805.02109.

Stanley, Isaac, Alex Glennie, and Madeleine Gabriel. 2018. 'Nesta: How Inclusive Is Innovation Policy? Insights from an International Comparison'. Nesta.

Tarasconi, Gianluca, and Carlo Menon. 2017. 'Matching Crunchbase with Patent Data'.

Tatli, Ahu, and Mustafa F. Özbilgin. 2012. 'An Emic Approach to Intersectional Study of Diversity at Work: A Bourdieuan Framing: Emic Approach to the Study of Diversity'. *International Journal of Management Reviews* 14 (2): 180–200. https://doi.org/10.1111/j.1468-2370.2011.00326.x.

Thomson, Emily. 2009. 'Do Ends Justify Means? Feminist Economics Perspectives on the Business Case for Gender Equality in the UK Labour Market'. *E-Cadernos CES*, no. 05 (September). https://doi.org/10.4000/eces.298.

Zehavi, Amos, and Dan Breznitz. 2017. 'Distribution Sensitive Innovation Policies: Conceptualization and Empirical Examples'. *Research Policy* 46 (1): 327–36.

https://doi.org/10.1016/j.respol.2016.11.007.