WARWICK
ECONOMICS

Exploration and Exploitation in US Technological Change

Vasco M. Carvalho, Mirko Draca, and Nikolas Kuhlen

(This paper also appears as CAGE Discussion paper 575)

August 2021 No: 1366

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

Exploration and Exploitation in US Technological Change[⋆]

Vasco M. Carvalho, Mirko Draca, and Nikolas Kuhlen[⋆⋆]

June 2019
This version: August 9, 2021

How do firms and inventors move through 'knowledge space' as they develop their innovations? We propose a method for tracking patterns of 'exploration and exploitation' in patenting behaviour in the US for the period since 1920. Our exploration measure is constructed from the text of patents and involves the use of 'Bayesian Surprise' to measure how different current patent-based innovations are from existing portfolios. Our results indicate that there are distinct 'life-cycle' patterns to firm and inventor exploration. Furthermore, exploration activity is more geographically concentrated than general patenting, but this concentration is centred outside the main hubs of patenting.

[⋆] We thank Tiago Cavalcanti, Petra Geraats, Bill Janeway, Alexei Onatski, Christopher Rauh, Mikhail Safronov, and Weilong Zhang for helpful comments and suggestions. We thank participants at the Barcelona Graduate School of Economics Machine Learning for Economics Summer Forum 2019, Turing Research Showcase 2020 at The Alan Turing Institute, Cambridge Applied Micro Seminar and Monash-Warwick-Zurich Text-as-Data Workshop for useful feedback and suggestions. Kuhlen gratefully acknowledges the financial support of The Alan Turing Institute under research award No. TU/C/000030.

[⋆⋆] Carvalho: University of Cambridge (vmpmdc2@cam.ac.uk); Draca: University of Warwick and CAGE Research Centre (m.draca@warwick.ac.uk); Kuhlen: University of Cambridge and The Alan Turing Institute (nk490@cam.ac.uk).

1 Introduction

Technological change and innovation are central to the process of economic growth but are difficult to measure. Following Griliches (1990), efforts to measure technological change and innovation can be summarised according to whether they involve either information on innovation outputs (for example: patents, scientific papers) and inputs (for example: R&D, employment counts of scientists and engineers) as proxy indicators, or are based on the residual information about factor usage that is represented by total factor productivity (TFP).

These approaches face clear challenges when it comes to capturing qualitative change in the range and conceptual basis of technologies over time, as well as the experimental nature of many technological investments. At a fundamental level, innovation choices involve a trade-off between exploration and exploitation. Specifically, a firm may shift between 'exploiting' a breakthrough by developing a given technology in more depth or dedicating more effort to searching and experimenting in a new technological domain. This latter process of search can be characterised as continuing 'exploration'. The trade-off between exploration and exploitation has been a prominent feature of behavioural theories of the firm, following, for example, Cyert and March (1963) and March (1991).

In parallel with this, there is also a longstanding literature on firm size and innovation focused on the Schumpeterian 'Mark I versus Mark II' debate about the role of large firms in innovation over time. A central question here has been whether larger firms inherently tend towards producing incremental rather than radical innovations (Cohen, 2010; Nicholas, 2015). Given the economies of scale that are associated with size, a shift towards incremental innovation is compatible with firms entering 'exploitation' phases in their growth.

A further literature has discussed scientific and artistic creativity over the individual life-cycle. Creativity is widely thought to peak between the ages of 30 and 40 across a number of domains (Dennis, 1956; Lehman, 1960; Galenson and Weinberg, 2000; Jones, Reedy, and Weinberg, 2014). Recent work studying this question in the context of US patenting (Kaltenberg, Jaffe, and Lachman, 2021) is in line with this, finding that patenting rates peak around the early 40s and that measures of the quality or importance of patenting decline with age.

In this paper, we follow the exploration versus exploitation perspective on innovation and outline new empirical measures that render tangible how a firm or inventor moves through their 'knowledge space'. We implement this approach across US firms, inventors and counties, which we refer to as 'units' of innovation. Our principal contribution is to construct a new empirical measure of unit-level innovation from the corpus of patent texts. The measure that we put forward is based on the changes in the 'text information' implicit in a unit's patent portfolio. As such, it is distinct from and complements existing measures of innovation that are based on inputs, outputs or TFP.

We use unsupervised learning methods to measure shifts in a unit's patenting activities, defined in terms of topics that correspond to probability-weighted word clusters. In short, we identify phases of exploration by measuring how a unit moves across the 'topic space' of its patents. Bigger jumps across the topic space are identified as phases of heightened exploration while stable years are indicative of phases of exploitation.

More specifically, to measure exploration, we first use Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) as a dimension reduction tool that allows us to describe patent texts in terms of a latent topic structure. Applying LDA to a unit-level patent corpus yields two main elements. These are, firstly, a set of endogenously generated knowledge topics and, secondly, a distribution of these topics over the set of a unit's patents.

We then use a 'Bayesian Surprise' (Itti and Baldi, 2009) measure to quantify the extent to which the patents of a given unit in a particular year contain a new mixture of topics compared to what came before. The Bayesian Surprise concept is grounded in information theory and results in a measure that is defined according to informational 'bits'. The concept has general applicability across social and natural science settings. For example, Itti and Baldi (2009) show that Bayesian Surprise captures real cognitive processes as it predicts what a subject shifts their attention and gaze towards. In our application, the unit-specific past topic distributions function as a prior, to be compared to the topic distribution in the current period. In this way, we use Bayesian Surprise to evaluate how exploratory a unit is at different points in time according to movements across its latent topic space.

We build on this further to construct a measure of 'successful' exploration by adopting the resonance measure proposed by Barron, Huang, Spang, and DeDeo (2018). In short, this measure hinges on how exploration in the current period is different relative to past and future exploration. A unit might move into a different area of its underlying topic space but may not stay in this area. This would be an example of 'unsuccessful' exploration. In contrast, successful exploration is defined as episodes where exploration in the current period is different to past exploration but similar to subsequent firm innovation activity. 'Successful' exploration is therefore an episode of exploration that 'sticks' and is manifested in a lasting change in a unit's underlying topic distribution.

Our empirical implementation of this approach uses a database built up from a match of US Patents and Trademark Office (USPTO) records on the abstracts of patents to information on firms, inventors and counties. This provides us with a long time period for studying the evolution of these units. For firms, we are able to measure exploration behaviour for the period since 1920 while for counties and inventors we study the periods since 1947 and 1976, respectively.

Findings. We implement our exploration measure at a range of levels with a specific focus on identifying developmental patterns in the progress of exploration. We first demonstrate our methodology with a case study of the International Business Machines (IBM) corporation, a firm that was central to the development of computing technology during the 20th century. This case study shows how IBM underwent a major transition from mechanical and analogue to digital technologies in a period centred on the 1950s. This transition is apparent from the basic word frequencies of IBM's patents across decades, the underlying topic structure of the firm's patent portfolio, and from the summary exploration measures that we calculate.

We next look at patterns of exploration across all available firms. Using a measure of 'cumulative exploration' (in effect, the integral of annual exploration flows) we are able to trace out developmental patterns in a firm's innovation behaviour. That is, there are clear phases of faster and slower exploration, including evidence of widespread 'S-shaped' diffusion-style trajectories.

Interestingly, the principal explanatory factor for these firm exploration trajectories is firm age. The correlation with firm age actually dominates as an explanatory variable when it is included alongside firm size variables and a patent stock measure. There is also a clear 'wedge' between the exploration-age profile of firms versus the firm size-age profile. Practically, this means that exploration tapers off with age faster than firm size, hinting again at a potential developmental pattern in firm behaviour. This is complemented by a pattern of sharply declining Research and Development (R&D) intensity in firm age. On average, the early years of a firm's life in the US data we examine seem to be dedicated to (relatively) more pronounced exploratory and R&D-intensive innovation.

In the final part of our analysis of firms we examine the association between our exploration measure and firm sales growth. This indicates that there is an association that is robust to industry trends and controls for the growth of patenting. Furthermore, the association also holds when controlling for age, indicating that the intensive margin of exploration across firms of the same age has explanatory power. Our measure of successful exploration also appears to be effective at identifying phases of exploration that are more strongly associated with sales growth than the 'general' measure of exploration.

Our next set of findings focuses specifically on the geographical distribution of ICT patenting and exploration across US counties. We observe that exploration is more geographically concentrated than actual patenting itself, but that there is a limited overlap in the concentration of patenting. That is, exploration is occurring away from the main hubs of patenting, with the top examples of this intensive 'periphery' exploration being counties where defense contracting firms have a strong presence. Overall, we find that the concentration of exploration was highest in the period between 1960-1980. The decline following 1980 then occurs alongside an increase in the concentration of ICT patenting itself, in this case towards classic innovation hubs such as Palo Alto.

The final application we look at relates to inventor age and exploration. As discussed, there is a broad literature that has found support for the idea that creativity and scientific productivity peak at middle age. Our findings are in line with this literature. We find that exploration peaks at around the age of 40 across a number of subsets of inventors – the full sample plus the 'superstars' in the top 1% and 0.1%. There are indications that the superstars defined in terms of the volume of patents produced go through 'waves' of exploration but a conventional, middle-aged peak holds for superstars identified according to average lifetime exploration. The life-cycle peaks in exploration are also substantial: inventors are around twice as exploratory at their peak than they are at other periods of life.

Related Literature. In addition to the work on firm growth, inventor life-cycles and economic geography that we have discussed this paper contributes to the emerging literature on using text-based information to measure innovation. Kelly, Papanikolaou, Seru, and Taddy (2018) construct a measure of 'breakthrough patents' using historical USPTO data and following a principle of 'backward importance'. That is, breakthrough patents are those that are amongst the first to feature n-gram phrases that became more common in later patents. Bussy and Geiecke (2020) follow the same intuition of comparing patent similarity across past and future periods but with an implementation focused on Latent Semantic Analysis (LSA) methods. The identification of new or fast-growing in patents is also at the centre of the contributions by Balsmeier, Assaf, Chesebro, Fierro, Johnson, et al. (2018), Bowen, Fresard, and Hoberg (2021) and Packalen and Bhattacharya (2015). Arts, Cassiman, and Gomez (2018) provide a deep discussion of the measurement of patent text similarity, with the additional element of introducing expert (human) validation to their basic framework.

Our main contribution to this literature is to provide a text-based measure of innovation that operates directly at the unit rather than patent level. That is, rather than identifying individual patents that are novel in their use of new and latterly important words we focus on the evolution of a firm, inventor or geographical area's overall patent portfolio. We are also unaware of any work on the economic modelling of innovation that has been rooted in the Bayesian Surprise concept, which has shown much utility in applications related to cognitive science (Itti and Baldi, 2009), cultural evolution (Barron et al., 2018), and the history of scientific thought (Murdock, Allen, and Dedeo, 2017).

The remainder is organised as follows. Section 2 introduces the methodology of our exploration measures. Section 3 describes the construction of our data set. Section 4 discusses the IBM case study. Section 5 applies our measures to the data and presents our main results. Section 6 concludes.

2 Measuring Exploration

To identify exploration and exploitation patterns, we first reduce the dimensionality of the data by describing the patent texts in terms of their latent topic structure. To this end, we rely on Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) – a hierarchical Bayesian model for discrete data.

In general, our approach can be summarised as follows. We start by aggregating the patent texts to documents at our desired level of analysis. This can be, for example, at the firm-year level or represent other units of interest such as geographical regions, inventors, industries, or technology classes. We then probabilistically represent the position of each unit either the latent topic space. The topic space can be constructed for the unit-specific sub-corpus or the entire corpus of documents. Changes in the topic shares can subsequently be measured using the concept of Bayesian surprise.

The rest of this section discusses the methodology of our exploration measures in greater detail. Section 2.1 describes Latent Dirichlet Allocation. Section 2.2 discusses Bayesian Surprise. This is followed by the definition of the measures and a discussion of their properties in Section 2.3.

2.1 Latent Dirichlet Allocation

LDA is a probabilistic topic model. The generative process described by LDA assumes that a document is constructed as a mixture of topics. As such, LDA belongs to the class of mixed-membership models that attach multiple rather than a single class to each observation.

For each document, the mixed-membership property is expressed in terms of a probability distribution over latent topics. The topics are defined as probability vectors over all words forming the vocabulary, that is, each entry represents the weight a topic assigns to the corresponding term. In this way, a topic is characterised by the probability mass it places on a set of words expressing a common theme. Note that a word can be used to represent multiple topics with different probabilities. Intuitively, in our application to patent texts, a topic represents a technology.

The advantage of LDA over other natural language processing techniques is that the generative model provides a complete probabilistic interpretation. This allows to empirically compute information-theoretic quantities based on the inferred probability distributions. Specifically, the topic distribution represents a source sending a signal – the stream of words forming the document.

To generate a set of observed documents, LDA is formally specified in terms of the following process:

1. For each document d:
   a. Draw topic proportions $\theta_d | \alpha \sim \mathrm{Dir}(\alpha)$.

b. For each word $w_{d,n}$:

   i. Draw assignment $z_{d,n}|\theta_d \sim \text{Mult}(\theta_d)$.

   ii. Draw word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.

where K specifies the number of topics, $\beta_{1:K}$ are the topic specific word distributions over the vocabulary, and α is a K-dimensional Dirichlet parameter. θ_d represents the topic proportions, z_d denotes the topic assignments, and w_d are the observed words for the d-th document.

For a given collection of documents, the inferential problem is to compute the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)},$$

where $\boldsymbol{\theta}$, $\boldsymbol{z}$, and $\boldsymbol{w}$ denote the corpus-level sets of the respective document parameters. This posterior distribution is intractable. There are several procedures to estimate the parameters including both sampling and approximation-based algorithms. Since the patent corpora we analyse are potentially very large, we appeal to variational methods to perform posterior inference. We outline the approximate posterior inference procedure in Appendix B.

Model Selection. LDA belongs to the class of unsupervised learning algorithms. As such, the fundamental parameter to be prespecified when applying LDA is the number of topics K. In particular, there is a trade-off between a smaller number of topics leading to better human interpretability and a larger number of topics improving statistical measures of model-fit (Chang, Boyd-Graber, Gerrish, Wang, and Blei, 2009).

In our application to firm patent texts, we estimate individual topic models for each firm in the data set. Hence, searching for the optimal number of topics for each firm corpus is computationally expensive. Additionally, our focus lies on computing a summary measure of changes in the topic distributions rather than interpreting individual topics. For these reasons, we employ the following heuristic to set the number of topics depending on the size of the firm corpus. If a corpus of documents is comprised of more than 100 patents we set the number of topics to 50, for more than 1,000 patents to 100, and for more than 10,000 patents to 150. For corpora consisting of fewer than 100 patents we use ten topics. When estimating common topic spaces in the case of our county-level and inventor-level analysis, we set the number of topics to 100. Our results are robust to choosing different numbers of topics.

2.2 Bayesian Surprise

The concept of Bayesian Surprise by Itti and Baldi (2009) is the second key ingredient to the definition of our exploration measures. On an abstract level, Bayesian

Surprise is a measure of how data affects an observer and is rooted in information theory and Bayesian decision theory. The underlying principles are as follows.

First, the presence of uncertainty is a necessary condition for surprise to exist. Second, surprise represents a relative deviation from an observer's expectations. For instance, an observer may experience varying amounts of surprise at different points in time for the same data. Third, in a Bayesian framework, uncertainty is represented by probabilities that capture subjective degrees of beliefs. As data is acquired, the beliefs are updated from prior beliefs to posterior beliefs using Bayes' Theorem.

Building on these principles, Itti and Baldi (2009) define Bayesian Surprise as the difference between an observer's prior and posterior beliefs. Thus, only data which substantially affect the observer's beliefs yields surprise. They note that this is independent of the informativeness of the observation as measured by Shannon entropy, that is, the general uncertainty around the random variable's outcome.

Formally, Bayesian Surprise is computed as the Kullback-Leibler (KL) divergence from a prior distribution q to posterior distribution p

$$D_{\mathrm{KL}}(p||q) \ = \ \sum_{i=1}^{N} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}.$$

Rewriting the above equation yields

$$D_{\mathrm{KL}}(p||q) \ = \ \sum_{i=1}^{N} p(x_i) \left[\log_2 p(x_i) - \log_2 q(x_i) \right],$$

that is, Bayesian Surprise is equivalent to the expectation of the logarithmic difference between the prior q and the posterior p where the expectation is taken with respect to p. Note that when using a logarithm with base two, it is measured in bits. Also note that Bayesian surprise is asymmetric but invariant with respect to reparameterisations due to relying on the KL divergence.

2.3 Exploration Measures

Exploration. Based on the above definition of LDA and Bayesian Surprise, we construct our exploration measure. In general, we measure exploration from the perspective of an observer learning about the new patents applications in a given year. The observer's prior belief is the cumulative average topic distribution up to year t. That is, the observer expects the same average topics as observed in the past – exploitation is the expected default behaviour. We then measure exploration as the surprise the observer experiences when upon learning the topic distribution in year t. Put differently, we measure exploration as the temporary deviation from the past topic mean. This allows us to distinguish between phases of exploration and exploitation.

Formally, similar to the study by Murdock, Allen, and Dedeo (2017) in a cognitive sciences context, we define exploration as

$$\eta_t \ := \ D_{\text{KL}}\left(\theta_t \middle|\middle| \bar{\theta}_{-t}\right),$$

where

$$\bar{\theta}_{-t} \ = \ \frac{1}{t-1}\sum_{j=1}^{t-1}\theta_j$$

denotes the average topic distribution up until year t. In our applications, the topic distribution θ_t in a given year t is based on the collection of all documents filed by the firm, an inventor or in a given county that year.

Properties. We now interpret the technical properties of our exploration measure. The mechanics are the same for all three levels of aggregation in our empirical analysis, that is, inventors, firms and counties. First, note that in the case where the observed unit's topic distribution is exactly the same as the past average topic distribution, our exploration measure is equal to zero. This corresponds to a year in which they exploit accumulated knowledge. In case it is different from the past average, our measure is greater than zero. This corresponds to a year in which they explore new topics. Additionally, note that based on the construction of using the past average as a prior, the first time an inventor, firm or county explores a new topic, our measure will be higher compared to a situation where they pick up a topic it has already worked on in the past. Hence, our measure can be interpreted as temporal novelty.

Second, as pointed out above, exploration is asymmetric due to relying on the KL divergence. As a result, it has the desirable property of attaching higher weights in situations where the share of a topic increases compared to the opposite situation where a firm works less on a specific topic compared to the past average. Therefore, our measure not only measures the difference between the current and past distributions but it also takes into account their order. This property naturally corresponds to the definition of an exploration measure.

Cumulative Exploration. In addition to the above exploration flow measure, we are also interested in characterising the life-cycle of a firm in terms of different phases of exploration. That is, we not only address the question of how surprised an observer is in a given year but we also analyse the accumulated surprise an observer has experienced following the patenting activities by the firm in the past. For this purpose, we define the cumulative exploration or 'exploration stock' in a given year t as the monotonically increasing function

$$H_t \ := \ \sum_{j=1}^{t}\eta_j.$$

Successful Exploration. The flow and stock exploration measures allow us to quantify exploration in terms of deviations from the past topic mean. They do not, however, distinguish between successful and unsuccessful exploration. To identify phases of successful exploration, we adopt the resonance measure proposed by Barron et al. (2018). Resonance modifies the exploration measure by including a term that captures the future impact of new technologies.

In our application, this allows us to quantify the surprise of the patent topics in a particular year compared to the patterns of previous years and subtract the difference to future topics. High surprise given the past as a prior represents the firm-level exploration of new topics. High surprise given the future as a prior indicates that the firm does not continue working on the same set of topics. Hence, by considering both the initial novelty of the patents filed in a given year and the similarity to future patents, successful exploration is conceptually related to an innovation measure.

Formally, the measure is defined as follows

$$\rho_w(t) \ := \ \frac{1}{w}\sum_{d=1}^{w}\big[\mathrm{KL}\big(\theta_t||\theta_{t-d}\big) - \mathrm{KL}\big(\theta_t||\theta_{t+d}\big)\big],$$

where w is the window size. To put this in words, the measure uses the KL divergence to compare the topic distribution of year t to year $t-d$. From this, it then subtracts the KL divergence between year t to year $t+d$. The differences are averaged over all years that fall into a predefined window of size w around year t. Hence, the first term in the resonance measure corresponds to the novelty of the patents in a given year, while the second term captures whether a firm works on these topics in the future.

The resulting mechanics can be summarised as follows. Resonance in a given year t is low if the technologies are similarly different from past and future technologies or very similar to both. As a result, the measure is either equal to or close to zero. Positive resonance corresponds to situations where the current technologies are different from the past average and similar to subsequent technologies. This surprise asymmetry can be interpreted as successful exploration. Note that by construction, the two terms in the resonance measure are not symmetric. This is because the second term uses the future topic distribution as a prior.

One obvious drawback of using resonance in our application is that we require future information. Therefore, while we are able to identify historic phases of successful exploration, it cannot be used in a predictive way but rather complements our analysis.

3 Data

This section describes the construction of our data set from several sources relating to the three levels of aggregation in our analysis, that is, inventors, firms and counties. For a more detailed description of our text pre-processing steps we refer the reader to Appendix A.

Patent Abstracts. Similar to the analysis in Bergeaud, Potiron, and Raimbault (2017), we rely on patent abstracts rather than the full texts. The abstract should include the most important words that characterise the invention. Furthermore, the patent abstract focuses on the invention itself rather than including, for example, legal text.[1]

We obtain the abstracts from two main sources. Firstly, Bergeaud, Potiron, and Raimbault (2017) provide a database of abstracts for four million granted patents covering the period from 1975-2014. This database is derived from the electronic text patent records published by the USPTO. Directly inputted electronic records are not available before 1975 so we draw on a second database from Iaria, Schwarz, and Waldinger (2018). Their database was assembled from Google Patents files that were originally built up by applying optical character recognition (OCR) tools to scans of the original pre-1975 patent texts.

Formal abstract sections in patent text only became standard from the late 1960s onwards so we construct 'pseudo-abstracts' for this earlier period by subsetting the first 250 words of the patent document. This obviously relies on the assumption that the first 250 words are an effective summary of the overall patent. Our basic approach for defining the text of the abstract is to extract the text that lies between the two headings 'Abstract of the Disclosure' and 'Background of the Invention'. If the second 'Background...' heading cannot be found we define the 150 words after 'Abstract of the Disclosure' as the abstract text.

As a cross-check we compare the pooled pre- and post-1975 data to the list of 3 million patents from 1963 to 1999 that are included in the NBER legacy data set (Hall, Jaffe, and Trajtenberg, 2001). This revealed a set of 305,314 missing patents not covered by our main two datasets, so we directly webscrape information on this missing set of patents from the USPTO website. The pooled dataset across the three datasets covers 7,183,108 million patents granted between 1920 and 2014. Within this total, 2,466,973 patents are represented by pseudo-abstracts.

Patent Citations and Technology Classes. Our main source of data for patent citations and technology classes is the 'Comprehensive Universe of U.S. Patents' (CUSP) database constructed by Berkes (2018). In a similar vein to the abstracts, the citations are taken directly from computerized records for the post-1976 period and extracted from the text for the years prior to this. Berkes (2018) parses text from the 'References Cited' sub-section for patents issued between 1947-1975 and looks

[1] While there tend to be differences in topic coherence when comparing topics based on full-text to abstract data when extracting topics from small document collections, for large document collections these differences are less significant (Syed and Spruit, 2018)

across whole body of patent texts for the pre-1947 era, focusing on keywords that suggest the quoting of explicit patent numbers.

A novel aspect of the USPTO technology class field is that the USPTO regularly updates and corrects these classifications. This means that patents can be categorised according to a consistent modern taxonomy of classes. The three main classification systems in use are the International Patent Class (IPC), the Cooperative Patent Classification CPC) and the US Patent Classes (USPC). Berkes (2018) collects the USPTO classifications as at the date of June 2016 and defines a main class based on the distribution of disaggregated 3-digit classes for the CPC and IPC, while a main class is directly identified by the USPC system.

Firm Outcomes. To connect our exploration measure to firm outcomes, we use the Compustat and CRSP databases. Compustat contains information on listed company accounts from 1950 onwards while CRSP provides us with much more limited information based around stock prices and market value back to 1925.

We use the match of patent numbers to the CRSP 'permno' identifier from Kogan, Papanikolaou, Seru, and Stoffman (2017) to connect the two sides of the data. The Kogan et al. (2017) data provides information on 7,536 firms that are matched to 1.9 million patents from 1868 - 2009, although the years before 1920 and after 2005 are sparse due to censoring. For simplicity, we only use firms with a unique mapping of permno to gvkey as found in the CRSP crosswalk file, leading to a sample of 6,544 firms matched to the patent data.

Final Firm Data Set. Our exploration measures depend on a 'rolling window' structure whereby current period t topic distributions are compared to past and future distributions. This creates the restriction of requiring at least 11 years of continuous data in order to calculate firm-level exploration. In turn, our main sample is therefore a subset of 1,830 unique firms who account for 1,861,219 patents in total.

We calculate our measure of firm age from the joint firm-patenting database. That is, we infer the 'birth year' of the firm as the minimum year by permno. This captures the first year that a firm appears either in the USPTO patenting data or in the CRSP and Compustat firm data. For example, if a firm has taken out patents before it lists on the stock market, we are able to infer its existence on that basis. We finally drop the data from 2004 onwards to adjust for censoring effects such as the drop-off in patenting due to the lag between application and granting.

Geographical Data. The construction of the data sample for our county-level analysis is based on data set described above. We combine this with information on the assignee county and United States Patent Classification (USPC) classes provided by Berkes (2018). We then merge in the classification of USPC patent classes into technological categories and sub-categories following Hall, Jaffe, and Trajtenberg (2001). We obtain the mapping for this from Acemoglu, Akcigit, and Kerr (2016).

Lastly, we combine the annual exploration measure the population counts for each county from Manson, Schroeder, Van Riper, Kugler, and Ruggles (2020). Since the official population numbers are only available every five years, we linearly interpolate the population growth for the remaining years.

Inventor Age Data. For our inventor-level analysis, we obtain individual inventor identifiers and birth years for patents granted between 1976 and 2018 from Kaltenberg, Jaffe, and Lachman (2021). Their inventor birth years are inferred from information about inventors (name and location) combined with age information from different publicly available online web directories. We first merge this data with our full patent abstract sample. We then calculate the inventor ages as the difference between the application year of a patent and the birth year of the inventor. The resulting sample contains 3,264,210 patent texts matched to 1,354,897 individual inventors.

4 Case Study: International Business Machines (IBM) Corporation

To demonstrate our methodology, we first develop of case study of a single, long-lived firm. Specifically, we focus on the International Business Machines (IBM) corporation . IBM first emerged as a single corporation in the early 1920s from the merger of several previous companies with histories that go back to the 1880s. The company also had a central role in the development of computing technology in the 20th century, making it a good general example of the process of technological change.

We start by investigating changes in the raw word frequencies. In particular, we compute the change in the shares of a single word stem (unigram) in the total frequency counts used in IBM patents. This is constructed as a panel of the top 500 words per year for IBM's patents. The first column in Table 1 shows the top words across all years measured in terms of the levels. Unsurprisingly, the word "data" has the largest overall share. The remaining columns show the fastest growing unigrams calculated as the change in the share of the word in the total frequency count of words used in IBM patents per decade from the 1930s to the 1990s.

The table illustrates the shift in IBM's technologies over time. The early periods show IBM's focus on analogue apparatuses such as punched-card machines evidenced the use of words such as "gear", "time" and "sheet" in the 1930s and "card", "machin", and "tape" in the 1940s. For example, IBM managed the administrative information for the 26 million employment records that needed to be kept as part of the New Deal's Social Security Act of 1935.

The 1950s mark the transition from punched-card storage to digital storage (Bradshaw and Schroeder, 2003). This shift is evidenced by increases in count frequencies of words such as "circuit", "magnet", "memori", "data", and "signal". The

Table 1. Fastest Growing Unigrams by Decade for IBM.

Overall		**1930s**		**1940s**		**1950s**	
Word	Share	Word	Change	Word	Change	Word	Change
data	2.59	mean	1.64	card	2.81	circuit	2.52
system	1.45	feed	0.85	machin	1.68	magnet	1.63
layer	1.26	select	0.61	tape	1.10	memori	1.38
first	1.23	new	0.58	perfor	0.97	data	1.19
devic	1.13	gear	0.58	electron	0.69	signal	0.94
circuit	1.02	sheet	0.55	number	0.61	input	0.90
signal	0.94	time	0.55	sens	0.56	puls	0.87
second	0.92	applic	0.47	column	0.47	line	0.77
memori	0.84	charact	0.46	digit	0.47	devic	0.76
control	0.76	invent	0.43	valu	0.46	binari	0.63
1960s		**1970s**		**1980s**		**1990s**	
Word	Change	Word	Change	Word	Change	Word	Change
surfac	0.73	silicon	0.85	data	1.18	user	0.73
cell	0.60	line	0.78	system	1.04	layer	0.59
metal	0.58	layer	0.72	imag	0.53	system	0.56
control	0.55	print	0.55	comput	0.52	first	0.40
substrat	0.54	address	0.52	first	0.49	one	0.37
code	0.50	data	0.52	document	0.44	content	0.36
error	0.46	chip	0.50	access	0.42	request	0.34
wave	0.35	region	0.50	user	0.38	method	0.32
member	0.34	generat	0.40	circuit	0.35	process	0.31
mean	0.34	ribbon	0.38	optic	0.34	inform	0.30

Notes: This table shows the fastest growing unigrams (single words) per decade. This is calculated as the change in the share of the word in the total frequency count of words used in IBM patents. We construct this from a panel of the top 500 words per year for IBM's patents. The first panel shows the top words across all years measure in terms of the levels rather than changes in share. The units are percentage points (for example: 1.64 is 1.64%).

1960s to 1990s are characterised by words such as "surfac", "silicon", "data" and "user", respectively, representing the consolidation of the personal computer and the beginning of the internet.

Note that the growth rates after the 1950s are significantly smaller in magnitude compared to the previous period indicating that IBM stopped exploring and creating radically different inventions during this time but rather slowly adopted new technologies. This coincides with the period that lead up to the 'near-death' of the company in the mid-1990s.

We now illustrate how these changes observed at the high-dimensional word frequency level translate to the lower-dimensional topic space. First, to be able to visualise the evolution of topic shares, we run a separate ten-topic LDA model for

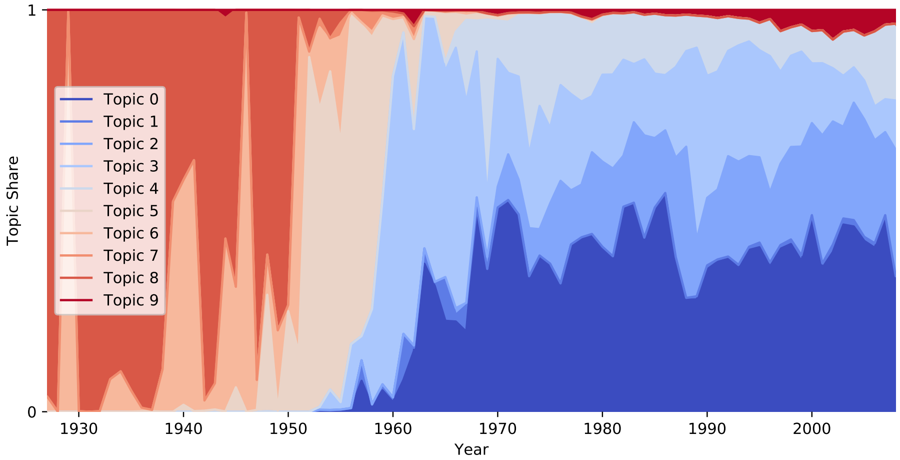


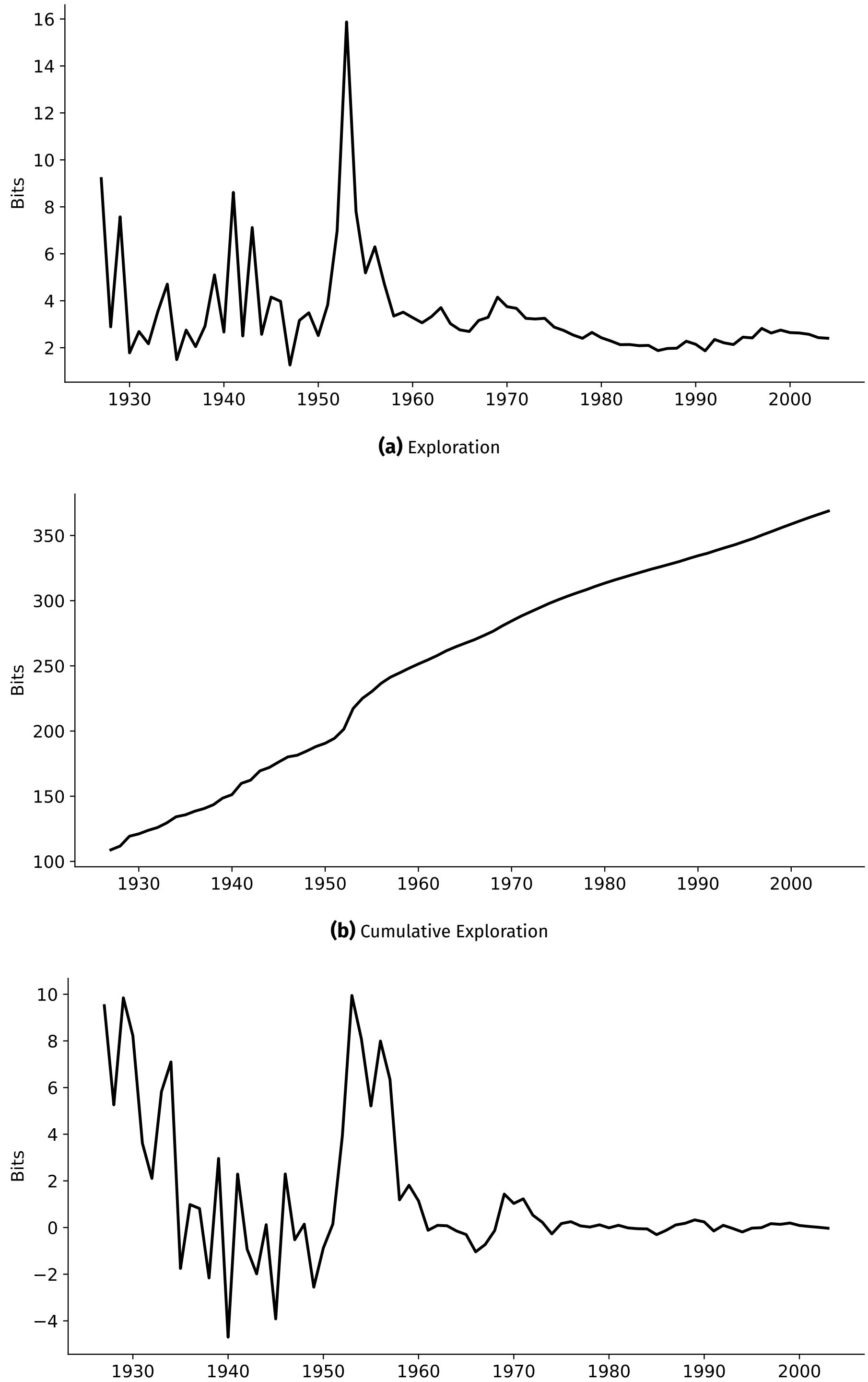
Figure 1. Evolution of Topic Shares for IBM.

Note: This figure illustrates the evolution of topic shares obtained from running a ten-topic LDA for IBM patents from 1927 to 2004.

IBM patents from 1927 to 2004 rather than the fully-fledged 150 topic specification. Figure 1 shows the evolution of the inferred topic shares over time. Most prominently, the graph illustrates the shift in the shares from analogue topics to digital topics in the 1950s. Furthermore, the periods before and after this transition are characterised by distinctive patterns. During the analogue era, IBM's topic shares are rather volatile implying that the attention given to individual topics is subject to rapid shifts. The digital period is marked by more equally distributed topic shares and generally less volatility.

Next, we show how our exploration measure summarises this information. Figure 2 shows the exploration, cumulative exploration and successful exploration time series for IBM from 1927 to 2004 based on the topics from the full 150-topics model. The exploration graph in Figure 2a exhibits clear phases of exploration and exploitation which correspond to the illustration of the topic share evolution for the ten-topic LDA model. Obviously, the largest spike in exploration corresponds to the aforementioned shift from analogue to digital technologies. IBM's early growth period up until the 1950s is characterised by higher exploration volatility capturing the radical shifts in topic attention described above. Starting from the 1960s, exploration is less volatile and smaller in magnitude which can be interpreted as a long phase of exploiting the previously developed technologies. Figure 2b visualises the corresponding accumulation of exploration over time. Naturally, the spike in the 1950s leads to a clear bump in cumulative exploration.

Lastly, Figure 2c displays the successful exploration as measured by resonance. The overall graph exhibits a very similar shape as the exploration series. The 1930s



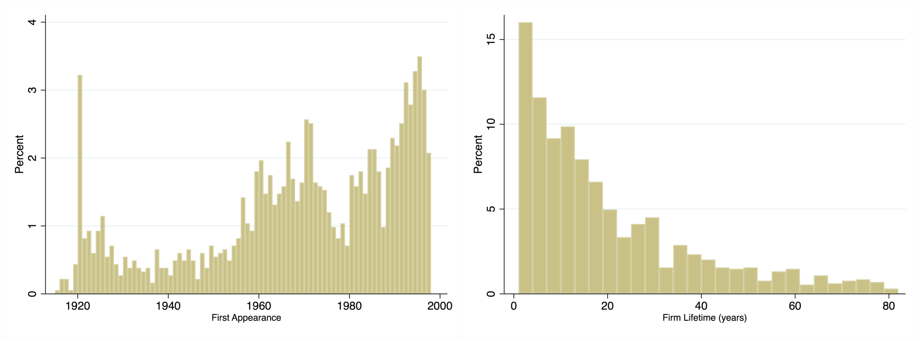
(a) Exploration

(b) Cumulative Exploration

(c) Successful Exploration

Figure 2. IBM's Exploration.

Notes: This figure shows the standard, cumulative and successful exploration series for IBM from 1927 to 2004.



(a) Entry Year of Firms

(b) Firm Lifetimes

Figure 3. Firm Age and 'Lifetimes'.

Notes: Figure 3a shows the 'entry year' of the 1,830 firms in our sample. This is defined as the first year a firm (i.e. a unique PERMNO) appears in either the CRSP, Compustat or matched USPTO data. The histogram bars are defined as 1-year intervals. Figure 3b shows distribution of firm 'lifetimes', defined as the number of times a firm appears across distinct years in the joint Compustat-CRSP-USPTO data. We calculate this on the cross-section of firms existing at or before 1990 in order to deal with right censoring (i.e. the fact that shorter-lived firms haven't played out their life-cycles). This represents 1,286 distinct firms, with an average lifetime of 19.9 years and median of 14.

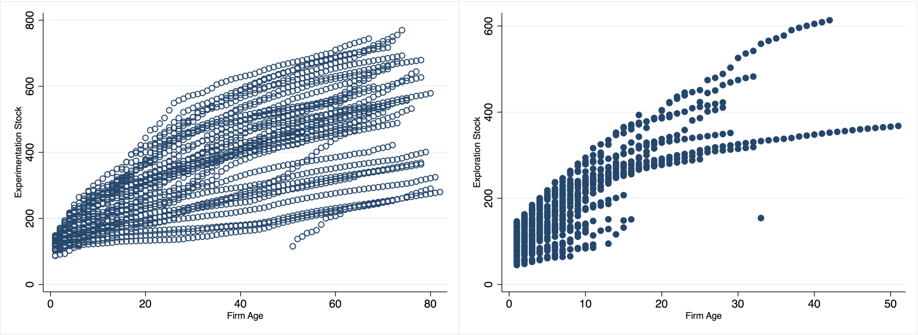
show a large spike in successful exploration. As before, the 1940s are characterised by high volatility, including negative spikes. That is, during this time IBM worked on topics that they dropped in future years. Note that 1940 marks the overall minimum of the series. As before, the 1950s show a large increase in successful exploration – the transition from analogue to digital storage. After a small positive bump in the 1970s, the graph stays flat around the zero line representing a long period without significant innovations having a lasting impact.

5 Empirical Results

This section applies our measure to the data set from the previous section and presents our main results. Section 5.1 investigates exploration patterns in firm behaviour and connects our measure to firm outcomes. Section 5.2 examines how exploration in ICT is distributed across counties. Section 5.3 investigates the relationship between exploration and inventor age.

5.1 Exploration, Firm Age, Firm Size, and Firm Growth

Firm Age and Lifetimes. We start our analysis by presenting some information on firm ages and 'lifetimes'. Figure 3a shows the distribution of firm birth years amongst unique firms in the cross-section. As discussed, this is calculated as the first year a firm appears in our joint USPTO-CRSP-Compustat database.



(a) Large Firms

(b) Semiconductor Industry

Figure 4. Exploration Stock and Firm Age Over Time.

Notes: Figure 4a shows the evolution of our 'experimentation stock' measure for firms that are aged 60 or more as of 2007 and are above the 95th percentile in the firm-level distribution of total cumulated patents (practically, 3,357 patents). N = 35 for the number of firms included. Average age of firms is 72.1 years. Figure 4b shows the evolution for firms in the semiconductor industry (SIC4=3674). N = 82 firms. Average age of firms is 10.9 years. Average cumulated number of patents per firm is 1,257.7. The SIC code assigned in Compustat from 1950 onwards is assigned for firms existing as part of the CRSP data pre-1950. Finally, note that the isolated 'dot' in Figure 4b is a firm with exactly 11 years of patenting and is therefore subject to the 'windowing' needed for the successful exploration measure.

Figure 3b then plots a histogram on firm 'lifetimes' in the cross-section. In the computation, we consider all of the unique firms that existed before 1991 and calculate the total number of years they are contained in our data. The conditioning of the data on 1990 and before helps to account for censoring – by definition those firms that have been born recently still need time for their commercial life-cycles to play out.

Firm Topics. Before examining exploration patterns, we briefly describe the process we rely on to construct firm-year documents for the LDA inferential procedure. In a first step, we combine all patents into a single document for each firm-year. We then normalise the length of this document to 100,000 words. The main reason for this is that the normalisation helps establishing comparability between years with different numbers of patent applications. The choice of 100,000 words is robust in the following sense. While shorter documents would introduce a noticeable bias to our exploration series, for document lengths above this number, our results do not change significantly. From these documents, we then infer the firm-level topic distributions which form the basis for our exploration measures and are used throughout this section.

Trends in Exploration. How does exploration evolve over the life-cycle of long-lived forms? Figure 4a displays the paths of the exploration measure for a sample of large, long-lived firms – aged 60 or older by the end of the sample and included in the top

Table 2. Relationship between Cumulative Exploration and Firm Age.

	(1) Baseline	(2) +SIC4	(3) +Mktcap	(4) +PatStock	(5) +Sales
age	12.03*** (0.533)	12.31*** (0.505)	11.89*** (0.511)	10.94*** (0.574)	11.73*** (0.576)
age2	-0.0645*** (0.00781)	-0.0644*** (0.00742)	-0.0633*** (0.00735)	-0.0562*** (0.00807)	-0.0607*** (0.00803)
log marketcap			7.156*** (1.746)		
log patstock				14.09*** (2.238)	
log sales					7.890*** (2.001)
R-sq	0.620	0.718	0.720	0.728	0.726
N	26,727	26,721	26,375	26,721	23,009

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of the cumulative exploration measure on firm age – age is the linear term while age2 is the quadratic. log marketcap is the logarithm of market capitalization, log patstock is the logarithm of the patent stock and log sales is the logarithm of sales. Year effects in all regressions, SIC4 fixed effects from Column (2) onwards.

5% of firms in terms of total patents. These paths show evidences of clearly defined trends at the firm-level, including indications of classic ‘S-shaped’ developmental behaviour.

We follow this up in Figure 4b by conditioning on all firms in the semi-conductor industry but relaxing any constraints on minimum firm age. This shows a pattern of dispersion whereby firms with higher exploration trajectories appearing to ‘break-away’ after surviving their first 10 years.

Next, we turn to regression models to further investigate this relationship. In particular, we aim at disentangling the question of how exploration varies with age and whether this relationship is conflated with firm size. We use the cumulative exploration measure or ‘exploration stock’ as the dependent variable. Table 2 shows the results for different specifications. The main message is that exploration is indeed parabolic in age and, interestingly, age explains exploration over and above any correlation with firm size. Specifically, Columns (3)-(5) control for market capitalisation, the firm patent stock and firm sales in succession with minimal effects on the coefficients of the two age variables. That is, age dominates as a stronger correlate of exploration, with this being clearly evident in the raw correlations. For

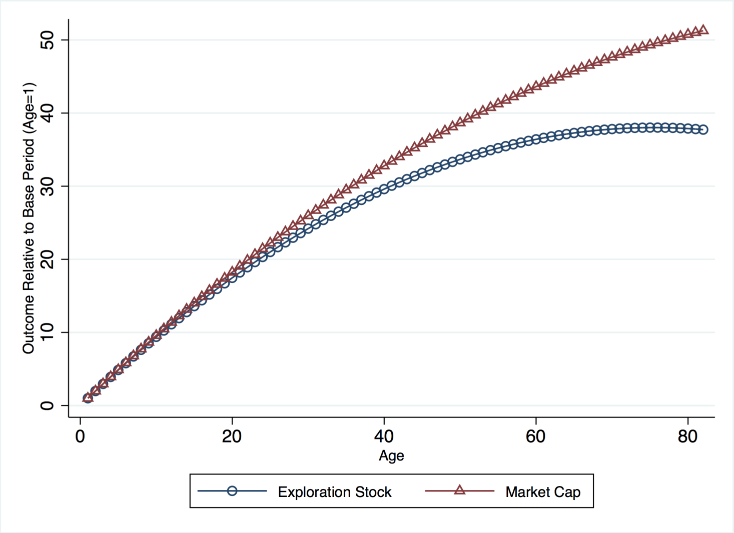


Figure 5. Gradients of Exploration Stock and Firm Size with Firm Age (All Firms).

Note: This figure shows the gradients of the exploration stock and firm size (defined as market cap) with firm age. This is defined as the predictions from a pooled cross-sectional regression of the outcomes on age and age squared with controls for year effects. N = 27,760 observations in the regression covering 1,795 distinct firms. The y-axis shows the level of the outcomes with respect to the age = 1 base period (i.e. we normalise with respect to initial values).

example, the age-exploration correlation is 0.83 compared to 0.38 for (log) market cap-exploration in the data underlying the Column (3) regression.

To summarise these relationships, we plot the age-firm size and age-exploration gradients in Figure 5. These gradients are the predictions from pooled cross-sectional regressions of the outcomes controlling for year effects. They show that exploration has a less steep slope with respect to age, that is, exploration tapers faster with age than with firm size. Theoretically, this is interesting insofar that it shows that firm growth continues after exploration has attenuated, hinting at the existence of major phases of exploitation activity amongst firms.

There is also a clear relationship between firm age and R&D intensity (defined as R&D expenditure divided by sales), which we plot in Figure 6. The graph shows that R&D intensity falls with age right up until age 40. Average R&D intensity in the early years of firm lifetimes is around 0.102 (i.e. R&D spending is 10.2% of sales) with a sample mean of 0.056. Again, this is prima facie evidence of intense exploratory behaviour earlier in firm life cycles.

Exploration and Firm Growth. We now connect our exploration measure to firm outcomes in a regression framework. We look at both the short-run dynamics of exploration and firm sales (effectively 1-year growth models) as well as medium-run relationships (5-year growth models). The basic model that we adopt is as follows:

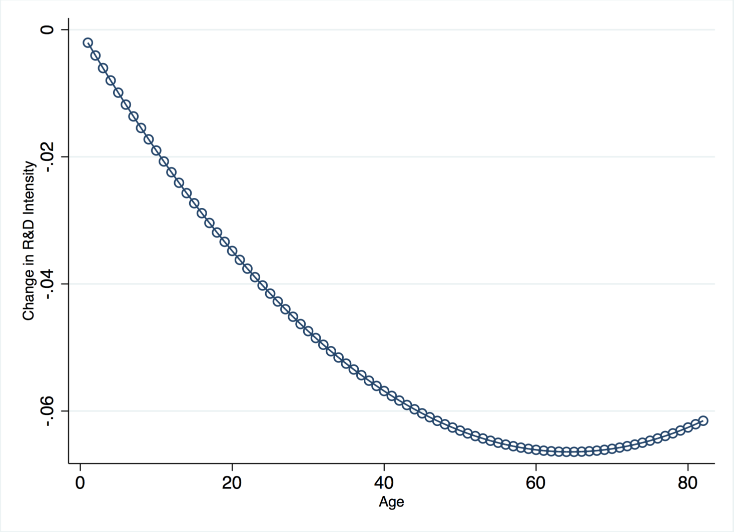


Figure 6. Change in R&D Intensity with Firm Age (All Available Firms).

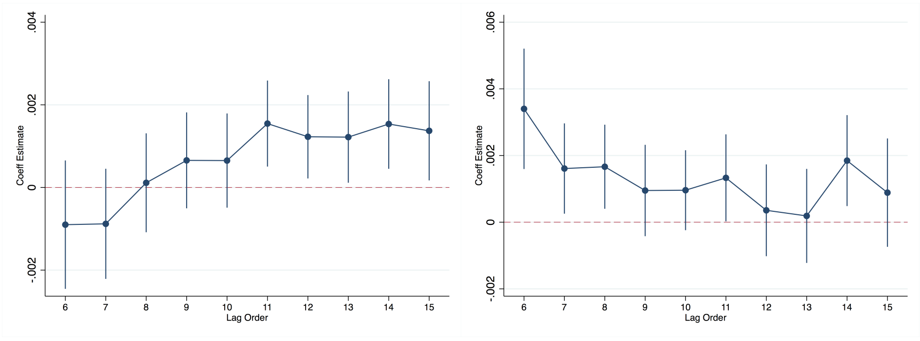
Note: This figure shows the gradient of firm R&D intensity (define as R&D expenditure over sales and firm age. This is defined as the predictions from a pooled cross-sectional regression of the R&D intensity on age and age squared with controls for year effects. N = 16,209 observations in the regression covering 1,467 distinct firms. The y-axis reports how R&D intensity changes with age. The mean R&D intensity across the sample is 0.056 while the mean starting value (i.e. at age=1) is 0.102.

$$\Delta_k \ln(\text{Sales})_{ijt} \;=\; \alpha + \sum_{L} \beta_{k-1} \text{KL}_{t-l} + \tau_t + \mu_j + \tau_{jt} + \varepsilon_{ijt}$$

where $\Delta_k \ln(\text{Sales})_{ijt}$ is the k-year change in firm i log sales measure in period t, KL_{t-l} is an l-period lagged exploration measure, τ_t are time effects, μ_j are industry effects, τ_{jt} are industry trends, and ε_{ijt} is an error term. We use different lag orders L to understand the dynamic relationship across specifications.

The main model that we focus on here is the 5-year changes model. This specification is useful for 'smoothing out' variation and reducing measurement error. In Figure 7 we present results for a specification that uses the 5-year change in (log) sales as the dependent variable and includes single-year exploration measures on the right-hand side. In effect, this is measuring the association between a 1-year shock in exploration at $(t-k)$ on a smoothed, 5-year measure of firm growth.

Figure 7(a) indicates that exploration has a medium-run association with sales growth. A positive association becomes evident at around the $(t-9)$ or $(t-10)$ lags, but is quite persistent once this point is reached. Note that this specification is run in changes and uses 'flow' measures in exploration so it is differencing out fixed unobservables at the firm-level. Figure 7(b) then runs a similar specification but uses successful exploration as the explanatory variables of interest. This shows a much sharper, short-run effect starting at the $(t-6)$ lag and is compatible with the idea that the successful exploration measure is better at picking out the most effective episodes of exploration.



(a) Ten years of Lagged Exploration

(b) Ten years of Lagged Successful Exploration

Figure 7. Five-Year Changes in Sales and Lagged Exploration.

Notes: This figure shows the estimates of a regression of the 5-year log change in firm sales on (simultaneous) lags of the general and successful exploration measures. Standard errors clustered by firm and 95% confidence intervals reported.

In Appendix C we present the results for a range of alternative specifications that relate sales to exploration. In Table C.1 we look at the relationship in terms of contemporaneous 1-year changes. This again shows a positive association that holds even after controlling for 4-digit industry trends, firm age and the change in the volume of patenting. The point estimate for successful exploration measure is also around three times higher than that for the standard exploration measure, confirming its effectiveness. We present the results of a similar 5-year changes specification in Table C.2. This differs from Figure 7 by using 5-year averages of exploration on the right-hand side and confirms the same patterns as the 1-year estimates.

What is the magnitude of this association? Our exploration measures are defined in terms of information 'bits'. Hence, for example, a 1-bit increase in exploration corresponds to an (approximate) 0.1 percent increase in sales in the specification in Column(3) of the upper panel in Table C.1. A 7.1 bit increase in exploration (which is equivalent to the standard deviation for this sample) then corresponds to a 0.9 percent increase in sales.

Recall here that the 'bits' are effectively measuring the extent of the *change in text information* in the firm-level patent portfolio. The regression specifications therefore show that firm sales performance is correlated with this change in text information over and above the quantity of patents being produced by firms.

5.2 The Geography of Exploration in ICT

In this section, we investigate how exploration is distributed across space. We focus on the specific context of patenting innovation in ICT, a key driver of U.S. innovation dynamics in post-war period. We are thus interested in understanding where exploration in ICT takes place, whether it is concentrated in particular exploration hubs and what, if any, are the dynamics of the geographical distribution of exploration in ICT.

We do this in the context of an increasing polarization of economic activity across space, at least partly driven by the rise of high-tech innovation hubs in the second half of the 20th century (Moretti, 2012; Moretti, 2019). Indeed, as shown by Andrews and Whalley (2021), after reaching a trough in the 1980s, the spatial concentration of patenting is today at an historical maximum, comparable to that observed in the mid-19th century. As their analysis documents, this is partly driven by the rise of ICT: by 2016, the commuting areas of San Jose (including much of Silicon Valley) and San Francisco, account for about nearly 20% of all U.S. patenting. Against this backdrop, we ask whether the spatial distribution of exploration in ICT simply reflects the patenting dominance of the familiar IT hubs or whether it is, instead, differentially concentrated.

The ICT subsample is then comprised of all patents belonging to category two ("Computers and Communications"). Given the focus on ICT, we further restrict the sample to patent applications made during the period from 1947 to 2007.

We then infer the topics by running LDA on the entire corpus of ICT patents aggregated at the county-year-level. This is followed by calculating the exploration measures based on the topic shares for each county. The advantage of this approach is that the topics are comparable across counties. In particular, for this exercise, we are interested in comparing the distribution and evolution of county-level exploration across the shared technology space rather than calculating within-county exploration. Hence, by using common topics, our resulting measures are not only comparable in terms their unit but also regarding the underlying topic structure.

Reflecting the highly spatially concentrated nature of patenting in ICT, the typical US county does not innovate in ICT: over the sixty-year period we consider, 2723 counties (out of a total of 3167) have zero patents, a further 285 counties patent only sparsely in ICT, with less than three patents per year on average, while the top 5% of counties account for 98% of all 452,889 ICT patents issued during this period. Henceforth we concentrate our analysis on this latter subset of counties accounting for the vast majority of ICT patenting.

Table 3 along with Figure 8 provide further confirmation of the spatial concentration of ICT patenting in the US post-war period. In particular, we compute, for each county, the total number of issued ICT patents as a share of the national grand total over the 1947-2007 period. Columns (1) to (3) of Table 3 rank the top ten counties while Figure 8a gives a heat map of its distribution across space. Consistent with our discussion above, the top ten counties account for nearly 90% of all ICT patenting during this 60 year period, with Santa Clara County alone (where Palo Alto is located) accounting for large 31% of all ICT patents and Westchester County (NY), where IBM headquarters are located, accounting for a further 16%. Also present in this top ten are the hubs of large metro areas (New York, Chicago's Cook County, Seattle's King County, Houston's Harris county, Los Angeles and New Jersey's Union

Table 3. Top Ten ICT Patenting and Exploration Counties.

(1) Rank	(2) County	(3) Share	(4) Rank	(5) County	(6) Share
1	Santa Clara County (CA)	31%	1	Madison County (AL)	27%
2	Westchester County (NY)	16%	2	Maricopa County (AZ)	14%
3	New York County (NY)	9%	3	Contra Costa County (CA)	9%
4	Cook County (IL)	7%	4	Alameda County (CA)	9%
5	King County (WA)	7%	5	Pima County (AZ)	9%
6	Middlesex County (MA)	6%	6	Marin County (CA)	5%
7	Harris County (TX)	5%	7	Riverside County (CA)	4%
8	Los Angeles County (CA)	5%	8	San Francisco County (CA)	4%
9	Union County (NJ)	4%	9	Orange County (CA)	3%
10	Dallas County (TX)	4%	10	San Diego County (CA)	3%

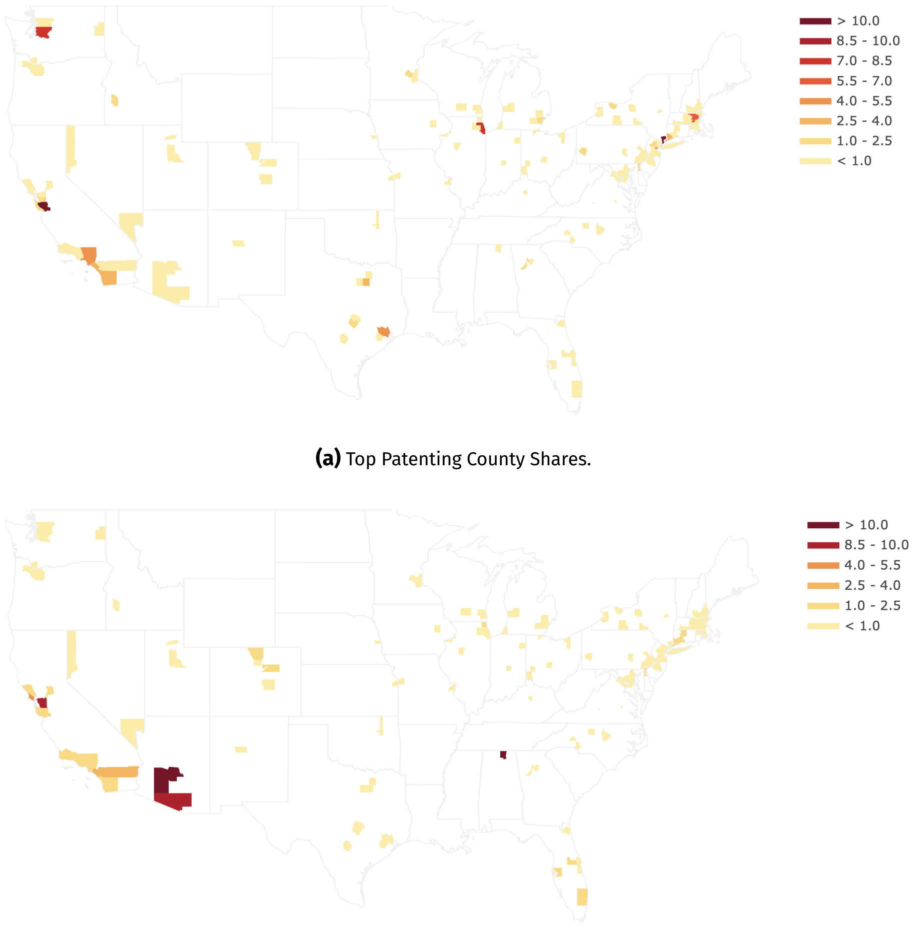
Notes: The table shows the top ten counties by shares of patenting (left) and exploration (right).

County) as well as Middlesex County (MA), where Cambridge is located.[2] The map visualises that the counties accounting for the remaining ten percent of patenting are spread across the country with the main areas located in the East and West.

Columns (4) to (6) in Table 3 show the ranking of the top ten counties with the highest total exploration over the sample period. The main observations from the table are as follows. First, there is no intersection between the previous group of top ten patenting counties and the top exploring counties. This indicates that the number of ICT patents does not necessarily capture the exploratory dimension of firm innovation behaviour. This is supported by an overall rank correlation of 0.02 for all counties. Second, exploration is more concentrated at the state-level compared to patenting. In particular, nine out of the top ten ICT exploration counties are in the West, with seven located in California. Third, the exception to this previous observation is the top county Madison County (AL). The county alone accounts for 27% of the total ICT exploration. Together with the second most exploratory county Maricopa County (AZ), the top two counties represent 41% of exploration. Figure 8b displays the corresponding map illustrating that exploration is highly concentrated in the top ten counties that represent 87% of the total ICT exploration. We also observe the general concentration in the West in contrast to patenting.

To get a better understanding of the firms that drive the patenting exploration in the top ten counties. Table 4 shows the top five patenting firms for each county in the top ten. Table 5 shows the top five exploring firms for each county in the top ten.

[2] These findings, both regarding the scale of concentration and the identity of the particular top locations, are consistent with the patterns documented in Andrews and Whalley (2021), albeit specialized here to ICT.



(a) Top Patenting County Shares.

(b) Top Exploration County Shares.

Figure 8. Top ICT Patenting and Exploration Counties.

Notes: The figure shows the total number of issued ICT patents as a share of the national grand total over the 1947-2007 period.

Focusing first on patenting, and in particular in the top patenting firms present in the very top three counties (which account for more than half of all patents issued over the entire period), we recognize that ICT patenting in these counties is - perhaps not surprisingly - dominated by well-recognized computer hardware component manufacturers, such as Intel, Sun, HP, Cisco, IBM (across two locations) or Hitachi as well as communications devices and services firms, such as ATT, or Phillips and an older cohort of firms in the same sector, such as ITT, RCA or Dictaphone.

Interestingly, and consistently with the limited overlap between top patenting and top exploration counties, the firms appearing as top explorers in the top exploration counties are in general distinct. For example, Madison county, responsible for more than a quarter of all ICT exploration over this sixty-year period, is a ma-

Table 4. Top ICT Patenting Firms.

Santa Clara County	Westchester County	New York County	Cook County	King County
Intel	Intl Business Machines	At&T	Motorola Solutions	Microsoft
Sun Microsystems	Hitachi	North American Philips	Boeing Co	Boeing Co
Hp	Texaco	Itt	Zenith Electronics	At&T Wireless Services
Cisco Systems	Dictaphone	Intl Business Machines	Gte	Amazon.Com
Advanced Micro Devices	Itt	Rca	At&T	Sundstrand
Middlesex County	Harris County	Los Angeles County	Union County	Dallas County
Raytheon Co	Hp	General Motors Co	Lucent Technologies	Texas Instruments
Digital Equipment	Compaq Computer	Northrop Grumman	At&T	Stmicroelectronics Nv
Emc/Ma	Litton Industries	Rockwell Automation	Alcatel-Lucent	I2 Technologies
Honeywell International	Exxon Mobil	Trw	Exxon Mobil	E-Systems
Gte	Halliburton Co	Directv	Agere Systems	Dallas Semiconductor

Notes: The table shows the top five patenting firms for the top ten counties shown in Table 3.

Table 5. Top ICT Exploring Firms.

Madison County	Maricopa County	Contra Costa County	Alameda County	Pima County
Intergraph	Honeywell International	Bio-Rad Laboratories	Network Equipment Tech	Burr-Brown
Avco	Honeywell	Chevron	Lam Research	Ventana Medical System
Motorola Solutions	General Electric Co	Systron-Donner	Exar	
Sci Systems	Motorola Solutions	Schlumberger	Eastman Kodak Co	
Adtran	Gte	Intraware	Sybase	

Marin County	Riverside County	San Francisco County	Orange County	San Diego County
Autodesk	Steris	Chevron	Western Digital	General Dynamics
Sonic Solutions	Toro Co	Macromedia	Smithkline Beckman	Cubic
L3Harris Technologies		Dolby Laboratories	Rockwell Automation	Titan
Inference -Cl A		Sharper Image	Emulex	Viasat
Fair Isaac		Schwab (Charles)	Qlogic	Oak Industries

Notes: The table shows the top five exploring firms for the top ten counties shown in Table 3.

jor aerospace and defense industry hub. The U.S. Space and Rocket Center, NASA's Marshall Space Flight Center, and the United States Army Aviation and Missile Command are all located in this county. Thus, Madison's top exploration location reflects the presence major contractors in the aerospace and defense sector, such as Intergraph (an early developer of geographical information systems for real time missile guidance purposes), the Aviation Corporation's Research Laboratory (Avco) or SCI Systems, a major electronic component manufacturer for the defense industry, as well as communications networks firms like Motorola and Adtran. The presence of major contractors to the defense industry extends to other top exploration locations beyond Madison county: Honeywell Aerospace and Honeywell International (in Maricopa, AZ, also a aerospace and defense hub), Systron-Donner (in Contra Costa, CA), L3Harris Tech (in Marin county), Rockwell Automation (Orange County, CA) or General Dynamics, Titon, Cubic or Viasat, all in San Diego County (CA), another major defense industry hub.

Finally, it's worth noting that beyond aerospace and defense, top explorer firms in top exploration counties reflect a diverse set of sectors, such as energy (e.g. Chevron, Schlumberger) or life sciences (e.g. Bio-Rad Laboratories, Ventana Medical, Smithkline Beecham, Steris Corp.) alongside perhaps more recognizable electronics components and devices or software firms (e.g. G.E., Autodesk, Dolby or Western Digital).

Overall, the analysis above suggests that the differential geographical distribution of ICT patenting relative to ICT exploration reflects the fact that whereas patenting is dominated by the location of electronics super-star patenting firms (such as IBM), ICT exploration reflects (*i*) innovation activities across a broader spectrum of sectors and, in particular, (*ii*) a sizeable contribution of the aerospace and defense industry, therefore tracking its geographical distribution.

The findings above suggests that, over our sample period, both ICT patenting and exploration are highly concentrated (albeit in different locations). A set of questions follow suit. Is ICT exploration more concentrated across space than ICT patenting? Are there differential dynamics of spatial concentration? Finally, how do we deal with the fact that top ranked counties according to either criteria appear to reflect very different sized counties? For example, for the year 2000, the population of Santa Clara (CA) county is close to 1.7 million while Madison County (AL) is close to 300,000. To address these questions, we follow the dartboard approach by Ellison and Glaeser (1997). The latter gives an intuitive null model to observed concentration patterns over space: that which would obtain if innovation – be it patenting or exploration – was randomly distributed across space with weights given by the population distribution across U.S. counties.

In particular, we use our data to compute the index for concentration C_t at time t as follows:

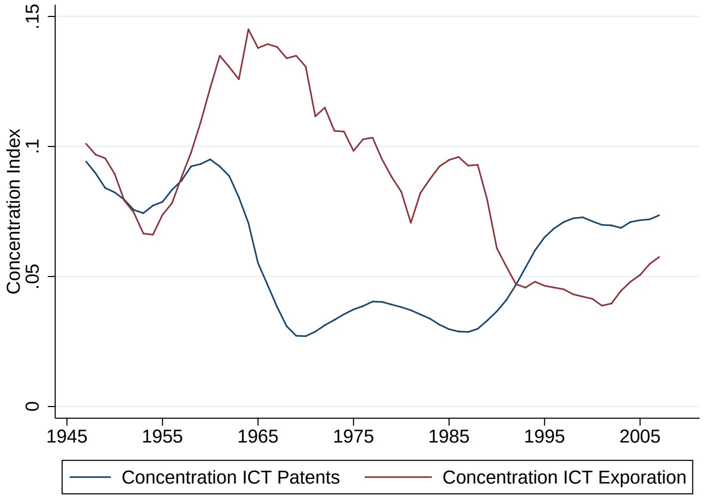


Figure 9. Spatial Concentration

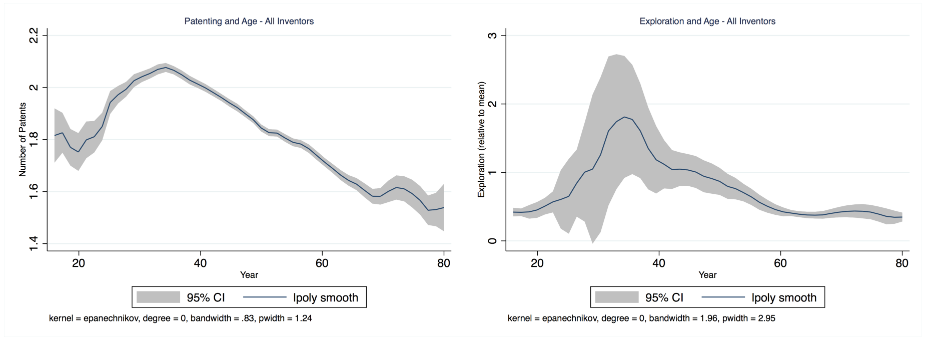
Note: This figure shows the spatial concentration of ICT patenting and exploration from 1947 to 2007 with a five-year moving average filter applied to each series.

$$C_t \;=\; \frac{\sum_{i=1}^{n}(\text{Innovation Share}_{it} - \text{Population Share}_{it})^2}{1 - \sum_{i=1}^{n}\text{Population Share}_{it}^2} \qquad (1)$$

where Innovation Share$_{it}$ is either the share of all exploration or all patents in ICT attributed to county i at time t. Whenever $C_t = 0$ this implies that each county innovation output is distributed according to its population share while if $C_t = 1$ all innovation in a given year t is attributed to a single county.[3]

Figure 9 displays the results, where we have applied a five-year moving average filter to each series in order to focus on lower frequency movements. First, note that both series display excess spatial concentration relative to the common benchmark, the spatial distribution of population. Second, the concentration of ICT patenting displays a market U-shape pattern, with spatial concentration falling by about 50% during the 70s and 80s (relative to the 50s and early 60s) and then rising again from the mid-90s onward. Further, these ICT patenting concentration dynamics are consistent with those reported by Andrews and Whalley (2021) for the entire population of US patents. Third, the average spatial concentration of exploration in ICT is higher than that of patenting (0.09 versus 0.06 sample averages,

[3] A related alternative would be to follow a dartboard approach of exploration relative to patenting. We would then be asking whether exploration is more concentrated relative to a case where exploration would be distributed across U.S. counties according to their respective ICT patenting shares. Not surprisingly, and anticipating results, this alternative approach yields similar findings to those presented above: exploration is more spatially concentrated than patenting but this excess concentration has declined over the decades.



(a) Patenting.

(b) Exploration.

Figure 10. Patenting and Exploration per Age for All Inventors.

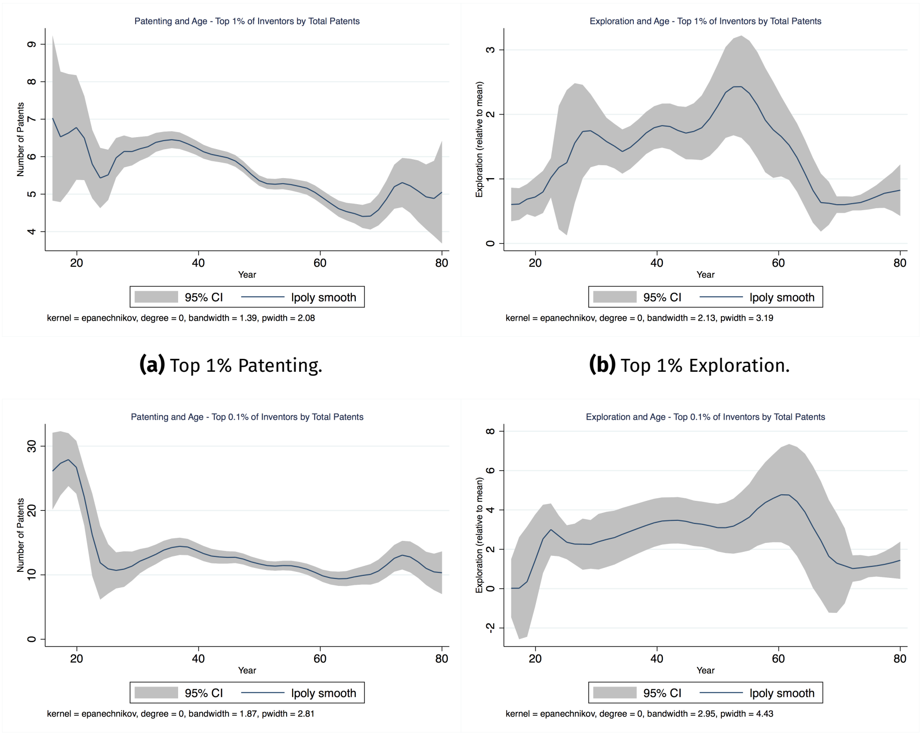
Note: This figure shows local polynomial regression plots for the sample of all N = 300,561 inventors with ages between 16 and 80. Patents are allocated in full to co-invented patents. The exploration measure is an index: exploration in 'bits' divided by the sample average of exploration.

respectively). Fourth, this is chiefly due to the different dynamics of the two time series. Thus, though they start at comparable levels of spatial concentration in the 50s, by the early 60s, when patenting concentration declines, exploration concentration increases (by about 50%) throughout that decade and, despite then initiating a trend decline, its excess concentration (relative to patenting) remains high throughout the 70s and 80s. By the same token, when we observe patenting concentration increasing again in the 90s, this is when we see exploration concentration declining below (that of patenting).

5.3 Exploration Over the Course of Life

In this section, we investigate the relationship between exploration and inventor age. Conceptually, our measure of exploration allows us to address the classic question of how scientific creativity varies with age. A broad range of research has suggested that creativity peaks in the age decades of the 30s and 40s (refs). Empirically, research on this topic has been obliged to use proxy measures of creativity such as patent or publication counts weighted by citations. In contrast, our exploration measure is designed to directly track how a researcher moves through 'knowledge space' over the course of their work.

We estimate inventor-level exploration by first estimating a 100-topic LDA model across all patent documents over all years. Exploration is then defined according to an inventor's topic shares for the portfolio of patents they produce in a given year. Hence, exploration in this context can be interpreted as measuring the shift in an



(a) Top 1% Patenting.

(b) Top 1% Exploration.

(c) Top 0.1% Patenting.

(d) Top 0.1% Exploration.

Figure 11. Patenting and Exploration per Age for Top Patenters.

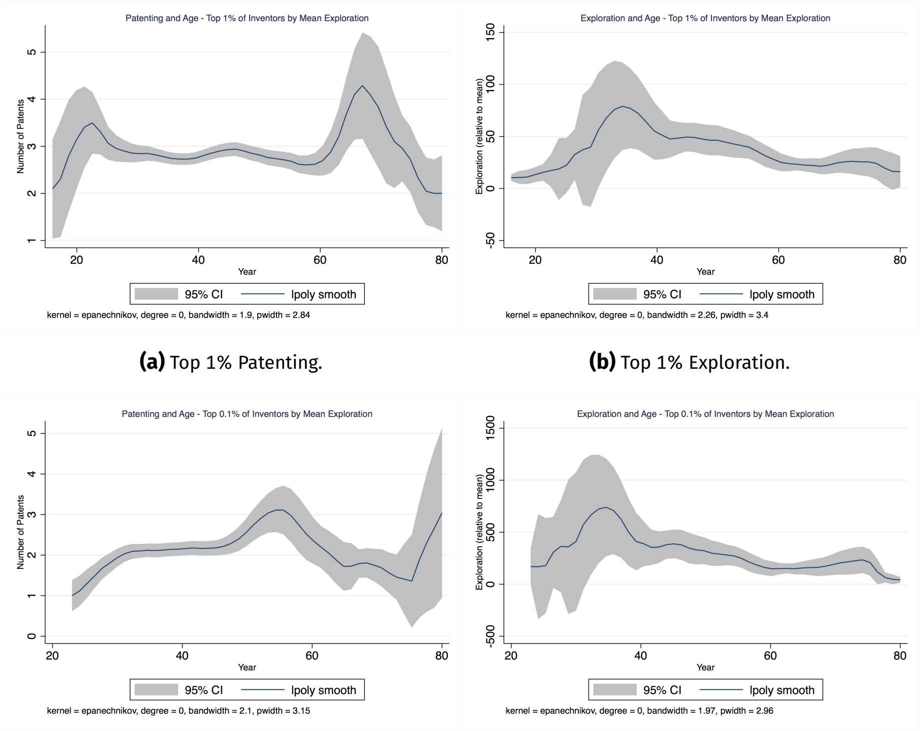
Note: This figure shows the results of local polynomial regressions for the samples of the top 0.1% and 1% patenting inventors.

inventor's pattern of specialisation across a set of topics defined at the level of the population corpus.[4]

Figure 10 shows the results of a local polynomial regression of outcomes on age for all inventors in the sample. In panel (a) we report the age profile of patenting – effectively patenting productivity over the life-cycle. The result here directly mirrors that of Kaltenberg, Jaffe, and Lachman (2021) – productivity in terms of patenting volume peaks around the age of 40 and then declines. Panel (b) then plots the profile for exploration, where we have normalised exploration according to the sample mean such that the y-axis can be interpreted as an index. This also shows a peak at around age 40. In this case, it is a steeper peak. Exploration is 2-3 times higher in the age 30-40 range than it is at other points in the life-cycle.

How does the exploration profile evolve for the most prolific inventors? We plot the age profiles for the top 1% of inventors by the number of total patents in panels

[4] Note that in contrast our firm-level analysis uses the firm-specific corpus to define the initial topic model, allowing for a 'within-firm' analysis of changing specialisation. We adopt the population-level corpus for inventors mainly for pragmatic (computational) reasons.



(a) Top 1% Patenting.

(b) Top 1% Exploration.

(c) Top 0.1% Patenting.

(d) Top 0.1% Exploration.

Figure 12. Patenting and Exploration per Age for Top Explorers.

Note: This figure shows the results of local polynomial regressions for the samples of the top 0.1% and 1% exploring inventors.

(a) and (b) of Figure 11 and then the top 0.1% in (c) and (d). This shows more variability in patenting productivity, with 'bursts' early and late in the life-cycle, but a high productivity mid-life phase is still evident. In terms of exploration, it should be first noted that the top 0.1% of inventors also tend to be more exploratory on average with indexed exploration levels of around 3.5-4 in mid-life compared to 1.5-2.0 for the full sample. Exploration also progresses in 'waves' across the life-cycle with a high level of exploration spread across the decades from the 30s to the late 50s.

We do an additional split by the top exploring inventors in Figure 12. That is, we calculate average exploration over the life-cycle and pick out the top 1% and top 0.1%. This results in a sample of inventors who produce an average of 2-3 patents per year. In this case, the pattern of exploration follows the more conventional pattern of peaking close to the age of 40 without any subsequent 'waves'. Arguably, what is most notable about this set of 'top explorers' is that a age profile is still evident even though these inventors have high baseline levels of exploration.

6 Conclusion

In this paper, we provide a new measure of unit-level exploration and exploitation. We empirically connect the measure to key questions in the literatures on firm growth, inventor life-cycles and the geography of innovation. We find evidence of exploration patterns in firm behaviour that are distinct from other potentially correlated aspects of firm performance, a mid-life peak in exploration for inventors, and evidence that exploration is geographical concentrated within the US but that this is coming from the 'periphery' rather than the main hubs of patenting.

The generalisability of our results faces a set of limitations. First, patent data is inherently biased towards a given unit's exploration activity that resulted in a patent application. Hence, while we rely on patents as an imperfect proxy for the total exploration activity, it is impossible to observe all innovation efforts. In addition, we only consider granted patents and thus exclude patents applications that were rejected. Second, in the case of firms and inventors our data set is subject to survivorship bias in the sense that we focus on the units with longer histories. Therefore, it is unclear how our results carry over to newer firms or inventors. Third, we currently do not take into account the effect of strategic interaction and renewal periods on patenting activity. Fourth, similar to most applications of natural language processing to a large, historic corpus there might be underlying changes in the patent language. However, since technical language typically faces less change compared to other written or spoken language, we deem this not to be too big of an issue.

We plan to develop the work in this paper in the following directions, with a strong focus on firms. Our first direction involves deepening the present analysis and further characterising the prevalence of exploration versus exploitation across the size and age distribution of firms. We also plan to aggregate our firm-level measures at the industry and economy level to explore a wider range of economic growth questions.

As a second direction, we will extend the breadth of our text-based measures of exploration. Our current measure focuses on the variance of exploration within a unit's life-cycle and have less explanatory power for studying how a unit's innovation behaviour is different from its peers. For example, an additional measure based on the Jensen-Shannon divergence would be better suited for quantifying unit deviations from group averages. There is also scope to complement our divergence measures with simpler metrics such as those based on how important, new words enter and diffuse through the patents text corpus.

Appendix A Data

This appendix describes the construction of our data set. Section A.1 discusses the general definition of patent abstracts. Section A.2 and Section A.3 describe our main sources of patent abstracts. Section A.4 describes the procedure we use to webscrape the remaining patents. Section A.5 discusses our text cleaning and pre-processing steps.

A.1 Patent Abstracts

We focus on utility patents filed at the United States Patent and Trademark Office (USPTO). More than 90 percent of UPSTO patents belong to the class of utility patents (Bergeaud, Potiron, and Raimbault, 2017). A utility patent provides intellectual property of an invention to its owner. As stated in Title 35 U.S. Code §101:

> *"Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title."*

The general conditions for patentability are novelty (35 U.S. Code §102) and non-obvious subject matter (35 U.S. Code §103). From 1860 to 1995, protection was granted for 17 years. Since 1995 the protection period has been increased to 20 years. According to PCT Rule 8 in the USPTO guidance, an abstract is supposed to be

> *"A summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention."*

A.2 Pre-1976 Patent Texts

We obtain the full patent text data for granted patents filed before 1975 from Iaria, Schwarz, and Waldinger (2018). The data set is constructed from digitalised versions of U.S. patents for grant years 1920 to 1979 from the web page of the USPTO. The patent texts were recovered using optical character recognition (OCR) scans and stored in plain text format. Note that the texts obtained from OCR may contain recognition errors introduced during the process of translating from image to text. These are typically caused by imperfections in the original scanned images. As pointed out by Kelly et al. (2018), going backward in time from 1976, the quality of OCR scans generally decreases due to a lower quality typesetting. The final data set is comprised of over 2.5 million patents with a total of more than 7.5 billion words.

Since our analysis focuses on patent abstracts, we extract the abstracts from the full texts where available using regular expressions. In particular, we consider the following three scenarios. First, if both section titles “abstract of the disclosure” and “background of the invention” can be found in the text, take the abstract as the text between the two titles. Second, in the case that the section title “background of the invention” is not contained in the full text but “abstract of the disclosure” and take the next 150 words as the abstract based on the UPSTO limit of 150 words for patent abstracts. Third, in cases where the abstract is not available, we extract the first 250 words of full text and use them as pseudo-abstracts.

A.3 Post-1976 Patent Texts

For patent abstracts of granted patents from 1976 to 2013 we rely on the MongoDB database created by Bergeaud, Potiron, and Raimbault (2017). They obtain the patent texts from USPTO bulk downloads. The total database consists of 4,666,365 utility patent abstracts.

A.4 Google Patents

When merging the above pre- and post-1976 data sets we find that they do not contain all patents granted when cross-checking against the list of three million patents from 1963 to 1999 in the NBER legacy data set (Hall, Jaffe, and Trajtenberg, 2001). We webscrape the text of patents that were not included in either of the two above sources from google patents.

A.5 Text Cleaning and Pre-Processing

After merging the three sources, we conduct a series of text cleaning and pre-processing steps. We begin by to converting terms into their linguistic roots. In particular, we extract word stems from the patent abstracts using the NLTK Snowball Stemmer. Note that the resulting word stems are not necessarily proper English words. We then use regular expressions to remove numbers and other non-alphabetic characters. Next, we remove occurrences of common stop words defined as terms that with little semantic content such as prepositions and pronouns appearing frequently in all texts.

This is followed by filtering out extremely rare or frequent words. Intuitively, frequent words are used in a majority of patents which in turn renders them uninformative with respect to a specific invention. At the same time, including rare words that are not integral to identifying a technology considerably increases the computational costs when applying our exploration and exploitation measures. For this purpose, we compute the term frequency–inverse document frequency (tf-idf) scores for each remaining keyword in each document. We use a sublinear (logarithmic) transformation to reduce the influence of extremely large or small scores. To reduce the size of the vocabulary, we remove all terms with a tf-idf score lower than

0.1. Finally, we eliminate all patents without any words left in their corpus after the previous removal step. The resulting data sample contains a total number of 277,019 distinct words.

Appendix B Approximate Inference

For a given collection of documents, the inferential problem is to compute the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{w}, \alpha, \beta) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} | \alpha, \beta)}{p(\boldsymbol{w} | \alpha, \beta)},$$

where $\boldsymbol{\theta}$, $\boldsymbol{z}$, and $\boldsymbol{w}$ denote the corpus-level sets of the respective document parameters. This posterior distribution is intractable. In the following, we outline the approximate posterior inference procedure. For a more detailed derivation, we refer the reader to Blei, Ng, and Jordan (2003).

The basic idea is to replace the above posterior by a fully factorised variational distribution

$$q(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_d q_\theta(\theta_d | \gamma_d) \prod_n q_z(z_{d,n} | \phi_{d,n}),$$

where the variational distribution of the topic proportions $\boldsymbol{\theta}$ is Dirichlet with parameter $\boldsymbol{\gamma}$ and the variational distribution of the topic assignments $\boldsymbol{z}$ is multinomial with parameter $\boldsymbol{\phi}$. This is followed by minimising the Kullback-Leibler (KL) divergence, or relative entropy, between the variational distribution $q(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\gamma}, \boldsymbol{\phi})$ and the true posterior $p(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{w}, \alpha, \beta)$. Note that minimising the KL divergence is equivalent to maximising the lower bound on the log likelihood of the observed documents $\log p(\boldsymbol{w} | \alpha, \beta)$ obtained from applying Jensen's inequality. This yields the variational updates

$$\begin{aligned} \phi_{d,n} &\propto \beta_{w_{d,n}} \exp(\mathbb{E}_q[\log(\theta_d) | \gamma_d]) \\ \gamma_d &= \alpha + \sum_n \phi_{d,n}. \end{aligned}$$

The variational updates have the following intuitive interpretation. The multinomial update corresponds to using Bayes' Theorem to obtain $p(z_n | w_n) \propto p(w_n | z_n) p(z_n)$. In the update equation, $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution. The update for the Dirichlet parameter is a posterior Dirichlet computed by adding the expected observation counts under the variational distribution $\mathbb{E}_q[z_n | \phi_n]$ to the pseudo-counts α (Blei, Ng, and Jordan, 2003). Using an Expectation Maximisation (EM) algorithm to maximise the variational lower bound yields the approximate empirical Bayes estimates. Specifically, the E-step consists of maximising the lower bound with respect to the variational paramters θ and γ. In the M-step, the bound is maximised with respect to the model parameters α and β. In our application to patent texts, we rely on the online variational Bayes implementation of LDA provided by the Gensim Python library.

Appendix C Additional Firm Figures

Table C.1. 1-year Changes in Sales and Exploration.

	(1) Baseline	(2) +SIC4	(3) +$\Delta_1$ln(PAT)$_t$	(4) +Age
Exploration$_{t-1}$	0.00160***	0.00138***	0.00132***	0.000786**
	(0.000238)	(0.000248)	(0.000248)	(0.000259)
$\Delta_1$ln(PAT)$_t$			0.00837***	0.00911***
			(0.00234)	(0.00233)
age				-0.00209***
				(0.000340)
age2				0.0000152***
				(0.00000340)
R-sq	0.048	0.067	0.067	0.070
N	22,738	22,732	22,732	22,732

	(1) Baseline	(2) +SIC4	(3) +$\Delta_1$ln(PAT)$_t$	(4) +Age
SuccessX$_{t-1}$	0.00273***	0.00244***	0.00226***	0.00216***
	(0.000390)	(0.000386)	(0.000384)	(0.000383)
$\Delta_1$ln(PAT)$_t$			0.0111***	0.0116***
			(0.00331)	(0.00331)
age				-0.00215***
				(0.000344)
age2				0.0000152***
				(0.00000335)
R-sq	0.055	0.079	0.080	0.084
N	19,835	19,826	19,826	19,826

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of the 1-year log change in firms sales $\Delta_1$ln(Sales)$_t$ on the 1-year lag of the general Exploration measure (top) and Successful Exploration (bottom). Year effects in all regressions, SIC4 fixed effects from col(2) onwards. $\Delta_1$ln(PAT)$_t$ is the 1-year change in log patent numbers log(1+PAT).

Table C.2. 5-year Changes in Sales and Average Lagged Exploration.

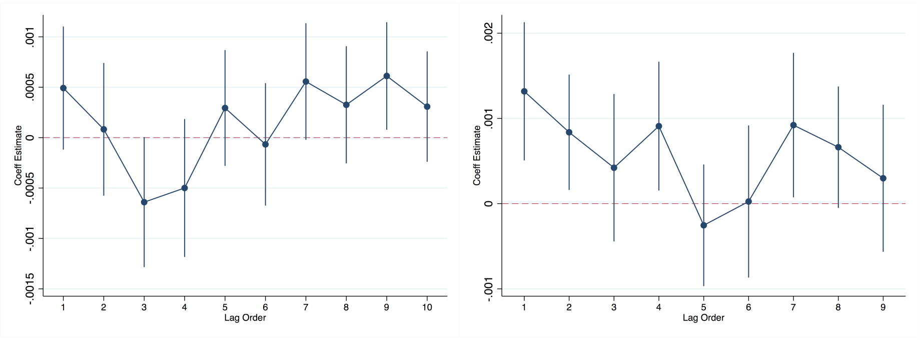
Panel (A)

	(1)	(2)	(3)	(4)
	1st-5-years	10-years	+$\Delta_5$ln(PAT)	all-available
Exploration$_{(t_6-t_{10})}$	0.00414	-0.000277	0.000848	0.00257
	(0.000212)	(0.00229)	(0.00224)	(0.00162)
Exploration$_{(t_{11}-t_{15})}$		0.00667***	0.00471*	0.00441**
		(0.00192)	(0.00198)	(0.00139)
$\Delta_5$ln(PAT)$_t$	0.0363***	0.0357***	0.0368***	0.0378***
	(0.00465)	(0.00466)	(0.00464)	(0.00446)
$\Delta_5$ln(PAT)$_{t-6}$			0.0223*	
			(0.00883)	
$\Delta_5$ln(PAT)$_{t-11}$			0.0189**	
			(0.00640)	
R-sq	0.210	0.213	0.316	0.134
N	10,865	10,865	10,865	20,719

Panel (B)

	(1)	(2)	(3)	(4)
	5-years	10-years	+$\Delta_5$ln(PAT)	all-available
SuccessX$_{(t_6-t_{10})}$	0.00991***	0.00976***	0.00907***	0.00976***
	(0.00228)	(0.00227)	(0.00260)	(0.00227)
SuccessX$_{(t_{11}-t_{15})}$		0.00486*	0.004	0.00486*
		(0.00246)	(0.00267)	(0.00246)
$\Delta_5$ln(PAT)$_t$	0.0421***	0.0421***	0.0423***	0.0421***
	(0.00511)	(0.00512)	(0.00507)	(0.00512)
$\Delta_5$ln(PAT)$_{t-6}$			0.00522	
			(0.0102)	
$\Delta_5$ln(PAT)$_{t-11}$			0.00630	
			(0.00898)	
R-sq	0.229	0.230	0.230	0.230
N	10,140	10,140	10,140	10,140

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of $\Delta_5$ln(Sales)$_t$ on general Exploration and Successful Exploration ('SuccessX'). The exploration measures are included as 5-year averages over the intervals of $(t_6 - t_{10})$ and $(t_{11} - t_{15})$. $\Delta_5$ln(PAT)$_t$ is the 5-year change in log patent numbers log(1+PAT) in period t. The 'All available' column in Panel (A) allows for taking averages in cases where all five 1-year lags are not defined. In Panel (B) this is the same as Column (2) since SuccessX requires continuous data in order to be defined.



(a) Ten years of Lagged Exploration

(b) Ten years of Lagged Successful Exploration

Figure C.1. One-Year Changes in Sales and Lagged Exploration.

Note: This figure shows the estimates of a regression of the 1-year log change in firm sales on (simultaneous) lags of the general and successful exploration measures. Stand errors clustered by firm and 95% confidence intervals reported.

References

Acemoglu, Daron, Ufuk Akcigit, and William R. Kerr. 2016. "Innovation network." *Proceedings of the National Academy of Sciences* 113 (41): 11483–88. DOI: 10.1073/pnas.1613559113. [12]

Andrews, Michael J., and Alexander Whalley. 2021. "150 Years of the Geography of Innovation." *Regional Science and Urban Economics*, (December): 103627. DOI: 10.1016/j.regsciurbeco.2020.103627. [23, 24, 29]

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez. 2018. "Text matching to measure patent similarity." *Strategic Management Journal* 39 (1): 62–84. DOI: 10.1002/smj.2699. [5]

Balsmeier, Benjamin, Mohamad Assaf, Tyler Chesebro, Gabe Fierro, Kevin Johnson, Scott Johnson, Guan Cheng Li, Sonja Lück, Doug O'Reagan, Bill Yeh, Guangzheng Zang, and Lee Fleming. 2018. "Machine learning and natural language processing on the patent corpus: Data, tools, and new measures." *Journal of Economics and Management Strategy* 27 (3): 535–53. DOI: 10.1111/jems.12259. [5]

Barron, Alexander T.J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, institutions, and innovation in the debates of the French Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (18): 4607–12. DOI: 10.1073/pnas.1717729115. arXiv: 1710.06867. [3, 5, 10]

Bergeaud, Antonin, Yoann Potiron, and Juste Raimbault. 2017. "Classifying patents based on their semantic content." *PLoS ONE* 12 (4): 1–22. DOI: 10.1371/journal.pone.0176310. arXiv: 1612.08504. [11, 34, 35]

Berkes, Enrico. 2018. "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts." [11, 12]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1. [3, 6, 37]

Bowen, Donald, Laurent Fresard, and Gerard Hoberg. 2021. "Rapidly Evolving Technologies and Startup Exits." *Working paper*, [5]

Bradshaw, R., and C. Schroeder. 2003. "Fifty years of IBM innovation with information storage on magnetic tape." *IBM Journal of Research and Development* 47 (4): 373–83. DOI: 10.1147/rd.474.0373. [13]

Bussy, Adrien, and Friedrich Geiecke. 2020. “A Geometry of Innovation.” *SSRN Electronic Journal,* (September 2019): 1–63. DOI: 10.2139/ssrn.3676831. [5]

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. “Reading tea leaves: How humans interpret topic models.” *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–96. [7]

Cohen, Wesley M. 2010. *Fifty years of empirical studies of innovative activity and performance.* Vol. 1, 1 C. Elsevier B.V., 129–213. DOI: 10.1016/S0169-7218(10)01004-X. [2]

Cyert, Richard M., and James G. March. 1963. *A Behavioral Theory of the Firm.* Prentice-Hall. [2]

Dennis, Wayne. 1956. “Age and Productivity among Scientists.” *Science* 123 (3200): 724–25. DOI: 10.1126/science.123.3200.724. [2]

Ellison, Glenn, and Edward L. Glaeser. 1997. “Geographic concentration in U.S. manufacturing industries: A dartboard approach.” *Journal of Political Economy* 105 (5): 889–927. DOI: 10.1086/262098. [28]

Galenson, D. W., and B. A. Weinberg. 2000. “Age and the quality of work: The case of modern American painters.” *Journal of Political Economy* 108 (4): 761–77. DOI: 10.1086/316099. [2]

Griliches, Zvi. 1990. “Patent Statistics as Economic Indicators: A Survey.” *Journal of Economic Literature* 28 (4): 1661–707. [2]

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. “The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools.” [11, 12, 35]

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. “Frontier Knowledge and Scientific Production.” *Quarterly Journal of Economics,* (June): 927–91. DOI: 10.1093/qje/qjx046.. [11, 34]

Itti, Laurent, and Pierre Baldi. 2009. “Bayesian surprise attracts human attention.” *Vision Research* 49 (10): 1295–306. DOI: 10.1016/j.visres.2008.09.007. [3, 5, 7, 8]

Jones, Benjamin, E.J. Reedy, and Bruce A. Weinberg. 2014. “Age and scientific genius.” URL: https://www.fatherly.com/wp-content/uploads/2016/03/w19866.pdf. [2]

Kaltenberg, Mary, Adam B Jaffe, and Margie E Lachman. 2021. “Invention and the Life Course: Age Differences in Patenting.” *National Bureau of Economic Research Working Paper Series* No. 28769: URL: http://www.nber.org/papers/w28769%7B%5C%%7D0Ahttp://www.nber.org/papers/w28769.pdf. [2, 13, 31]

Kelly, Bryan T., Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2018. “Measuring Technological Innovation over the Long Run.” *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3279254. [5, 34]

Kogan, Papanikolaou, Seru, and Stoffman. 2017. “Technological innovation, resource allocation, and growth.” *Quarterly Journal of Economics,* (November): 665–712. DOI: 10.1093/qje/qjw040.Advance. [12]

Lehman, H. C. 1960. “The age decrement in outstanding scientific creativity.” *American Psychologist* 15 (2): 128–34. [2]

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2020. “IPUMS National Historical Geographic Information System: Version 15.0.” [13]

March, J.G. 1991. “Exploration and exploitation in organizational learning.” *Organization Science* 2 (1): 71–87. DOI: 10.1287/orsc.2.1.71. arXiv: z0009. [2]

Moretti, Enrico. 2012. *The new geography of jobs.* New York: Houghton Mifflin Harcourt Publishing Company. [23]

Moretti, Enrico. 2019. “The effect of high-tech clusters on the productivity of top inventors.” *NBER Working Paper No. 26270*, [23]

Murdock, Jaimie, Colin Allen, and Simon Dedeo. 2017. “Exploration and exploitation of Victorian science in Darwin’s reading notebooks.” *Cognition* 159: 117–26. DOI: 10.1016/j.cognition.2016.11.012. [5, 9]

Nicholas, Tom. 2015. “Scale and Innovation During Two U.S. Breakthrough Eras Scale and Innovation During Two U.S. Breakthrough Eras.” [2]

Packalen, Mikko, and Jay Bhattacharya. 2015. “New Ideas in Invention.” *Working Paper*, DOI: 10.3386/w20922. [5]

Syed, Shaheen, and Marco Spruit. 2018. “Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation.” *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017* 2018-Janua: 165–74. DOI: 10.1109/DSAA.2017.61. [11]