Teams and Text: Collaborative Innovation in the Knowledge Space

Joseph Emmens[‡]

September 25, 2024

Click here for the latest version

Abstract

This paper proposes a novel framework that integrates inventor teams and the patents they produce together, building a spatial mapping called the *knowledge space*. I use this framework to study how the depth of prior work in a research field determines two innovation outcomes. These are the ability of a team to spark new and successful research fields, and their capacity to target innovation at specific objectives. For targeted innovations, I examine three directions: automating production, mitigating climate change and reducing cancer risks. I infer the knowledge space from patent texts using a Bayesian Natural Language Processing model. I then combine this with data on premature inventor deaths to provide quasi-random variation in team composition. A team boosts their likelihood of producing a breakthrough by reducing their size and moving into less populated areas of the knowledge space. However, when targeting a specific outcome, they benefit from adding a new member who moves them into established fields with many prior examples to build upon.

Keywords: Teams, Innovation, Patents, Topic Modelling

JEL: O30 O31 O32 O34 C50 C55 D83 L26

[‡]IAE CSIC, Universidad Autònoma de Barcelona and The Barcelona School of Economics. Thank you to my supervisor Hannes Mueller and to Christian Fons Rosen for all their valuable feedback. I would also like to thank Pau Milan, Inés Macho Stadler, David Perez Castrillo. All errors are my own. I gratefully acknowledge the Spanish Agencia Estatal de Investigación (MCIN/ AEI /10.13039/501100011033) through grant PID2020-114251GB-I00. The python code and do files which provide a replication of the estimation method can be found on my GitHub. Website: www.josephemmens.com

1 Introduction

Research teams are the driving force behind advances in science and technology. Organising inventors into effective teams is essential for creating new research fields and solving society's biggest challenges. Endogenous growth theory states that teams produce new technologies by recombining the existing knowledge of team members. This remains the central paradigm for innovation in economics. Yet, the collaborative knowledge production process remains a black box.

I propose an innovation production process that integrates inventor teams and the patents they produce together, building a spatial mapping referred to as the knowledge space. I use this framework to examine how prior work determines team output. Are teams more or less likely to produce breakthrough innovations when they have a substantial body of prior research to build on? Similarly, does extensive knowledge on a specific type of innovation increase the likelihood of a team patenting toward that objective? Gaining insights into this process can inform both public and private sector R&D policies. However, to do so, we require a model of teamwork that can measure how much and what type of prior work a team builds on. By locating inventor teams and patents in a unified space, I can evaluate the depth and characteristics of work in a team's research field.

This knowledge space is inferred from patent text data provided by the USPTO (PatentsView 2024) using a Bayesian Natural Language Processing model. This process uncovers important latent features, such as the distribution of knowledge among inventors and the contribution each individual makes to their team's innovations. I integrate this framework with data on premature inventor deaths provided by Kaltenberg, A. Jaffe, and M. E. Lachman 2021 to provide quasi-random variation in team composition. I examine two outcome variables: (1) whether a team's patent sparks a new research field (Kelly et al. 2021), and (2) whether their patents are aimed at a specific direction. For the second category, I focus on three key areas: automation, climate change mitigation, and cancer risk reduction (Mann and Püttmann 2023; USPTO 2024). I develop an empirical strategy around a set of continuous treatment TWFE models. The treatment captures the change in the quantity and type of prior work a team builds on following the premature death of a team member.

I find opposing effects for each of the two innovation outcomes. I show that teams produce fewer breakthrough patents as their local knowledge field becomes

more densely populated. Notably, a team's first patent is the most likely to be a breakthrough. This suggests a hit-or-miss innovation environment where team impact hinges on the success of their initial idea. For a group of inventors to produce a new scientific breakthrough, they often need to change research fields. I find that, following the premature death of a team member, the probability of the remaining inventors producing a breakthrough increases if the team shifts into a sparser area of the knowledge space.

In contrast to breakthrough innovation, I find that the likelihood of producing a patent targeting a specific direction increases with the number of local prior examples. Given this, teams that lose a member and subsequently move away from the relevant area of the knowledge space experience a significant decrease in the probability of patenting in that direction. I validate the robustness of this result by demonstrating that the effect reverses when a team adds a member instead of losing one.

This paper contributes a flexible but holistic model of teamwork to the literature. Prior literature relied on citation networks to track knowledge over time (Adam B Jaffe 1986; Hall, Trajtenberg, and Adam B. Jaffe 2001). While this provides a good representation of features of the patents, citations develop endogenously and capturing the emergence of a new field is challenging. A key contribution of this paper is the ability to disentangle who contributed which section of a patent. Previously, all citations or technology classes were allocated to each team member. The model presented here allows for far richer set of potential outcomes by allowing for the division of labour.

I begin by defining the knowledge space theoretically. This space consists of a fixed set of knowledge classes. Each class represents a specific domain of expertise, such as computer science, biology, or graphic design. Different combinations of these knowledge classes lead to different innovations. The knowledge space is modelled as a probability simplex across these classes, meaning each point in the space represents a unique combination of knowledge classes. Both inventors and patents are characterised by their position in this space. By using a simplex, this approach naturally incorporates a spatial concept by embedding a notion of distance.

I define a local knowledge field for each patent as the area within a fixed radius, where patents in this field share similar knowledge content. The spatial dimension of this model offers a significant contribution by intuitively capture when and where important fields emerge. I extend this concept to teams, where the area of the knowl-

edge space they cover shapes the type of innovation they produce. By co-locating inventors, teams and patents in one space I can examine how the quantity of prior work determines team innovation outcomes.

Innovation is modelled as a random process in which team members collaborate to create a new patent. Each inventor contributes knowledge according to their role within the team. Consequently, the team's output is determined probabilistically by the composition of its members, and the model derives a measure for an expected patent outcome. Allowing inventors to contribute differently to team output is fundamental to defining the team local knowledge field. This allows me to measure how much and what type of prior work does a team build upon.

I build on a method of Natural Language Processing (NLP) to empirically approximate the knowledge space using patent texts. Patents have been a valuable proxy of innovation for decades, and this paper forms part of a growing literature making use of the depth of knowledge contained in their texts. Through a hierarchical Bayesian model I infer who contributed which section of a patent text. Over each inventor's entire patenting history the model learns their individual knowledge distribution. If an inventor has a long history of producing AI patents and appears on a patent for a self-driving car with an inventor with a background in transport, the model can distinguish between their contributions. It identifies who provided the knowledge on automation and who contributed to the engine structure.

I validate this novel space along various dimensions. I classify breakthrough patents as those that experience the largest growth in the number of patents within their local knowledge field, following their publication. Patents that I identify as breakthroughs introduce 8.67% more new words and 47.6% more new combinations of two existing words, which are subsequently reused by future patents[1]. This evidence supports the central claim of this paper: breakthrough innovations are indeed driven by the recombination of existing ideas. I validate the contribution weights with the following reasoning. If one team member contributed significantly more to a patent, the technology classes of their past patents should provide more information when predicting the technology class of the current patent. When the gap between their contributions is large, the lead inventor's patenting history provides, on average, 14% more information. This difference disappears as their contributions become more equal.

[1]Using data kindly provided online by Arts, Hou, and Gomez 2021.

I leverage the mathematical foundations of the knowledge space, combined with economic theory, to develop two hypotheses. Endogenous growth is the central paradigm behind innovation, and this paper takes the recombinant growth model of Weitzman 1998 as a conceptual foundation. Specifically, I predict that prior work has opposing effects on team output. The impact depends on whether the team's objective is to create entirely new research fields or to redirect existing ones.

Prior work lowers the cost of innovation when a team focuses on ensuring their innovation meets a specific purpose (Grossman and Helpman 1991). If you are the first to design a smart version of a household object, it is challenging to transfer knowledge from computer science into your field. Following on from prior work reduces your costs of combining those knowledge fields, and thus increases the probability of you doing so.

When looking to produce a new research field, the existence of prior work is a barrier. Both mechanically by not being the first to develop the object, but also by defining paradigms that guide future work. This speaks to the literature on the burden of knowledge (B. Jones 2009). As the frontier expands endogenously, inventors now must invest more to reach that frontier. At the aggregate level, as the knowledge space fills up, breakthrough ideas get harder to find (Bloom et al. 2020).

These findings have key policy implications. To promote growth, policymakers should spread research funding across a wider range of fields. Concentrating resources in established areas often limits breakthroughs. However, policymakers are increasingly focused on guiding innovation. The results suggest encouraging collaboration across different fields. Firms looking to shift their focus can hire inventors from industries that have achieved similar goals. This brings valuable knowledge into their teams and facilitates cross-field learning.

Related Literature The first literature that this paper contributes to is on the importance of teams within science and technology. It is now taken as standard that teams are the principal producers of innovation (Wuchty, B. F. Jones, and Uzzi 2007). A range of reduced form papers have looked to describe team composition and its role in explaining innovation outcomes, where team size, hierarchy and diversity are key determinants (Xu, Wu, and Evans 2013; Wu, D. Wang, and Evans 2019; Uzzi et al. 2013). I contribute to this literature a holistic but parsimonious model of teamwork by extending the concept of mapping innovation (Fleming and Sorenson

2004) to the inventor, team and patent level. The spatial dimension allows me to measure various dimensions of team composition consistently, and understand their effect on the impact and direction of team innovations.

There has been significant interest in explaining complementarities between team members' knowledge and skills, where sorting into teams can explain aggregate innovation rates (Herkenhoff et al. 2024; Freund 2022; Pearce 2022; Boerma, Tsyvinski, and Zimin 2021). This literature has made important contributions to understand how teams create value, however to study how teams create new research fields and re-direct existing ones, a new framework was required. This paper proposes an alternative knowledge production process to the dominant CES production function which remains in keeping with the long literature on endogenous growth. This paper presents a model of how team members share their knowledge to innovate collaboratively, which is taken to the data on patent texts.

I contribute specifically to a smaller empirical literature which looks to disentangle individual contributions to team projects. A key development of the team production process presented here is allowing for inventors to contribute non-uniform shares to the knowledge contained in the innovation. Given the rising importance of teamwork, developing empirical methods to decode how teams combine individuals is key to understanding their production process (Ahmadpoor and B. F. Jones 2019). There is a small but important literature using highly specific case studies in which individual inputs are observed such as sports and experiments (Devereux 2018; Kahane, Longley, and Simmons 2013; Weidmann and Deming 2021). Alternatively, to identify the marginal contribution of a team member to the quality of team output Bonhomme 2022 employs a fixed effect model, but again is limited to relatively small team sizes.

Finally, the third literature I contribute to is the use of natural langauge processing models within social sceince studies. The ability to decode individual output lies in using high dimensional text data. Text analysis as a whole is booming, and following the seminal paper of Hansen, McMahon, and Prat 2018, LDAs have grown in popularity within economics. The closest paper to mine is the forthcoming study from Teodoridis, Lu, and Furman 2022. They develop a version of an LDA, a Hierarchical Dirichlet Process (HDP) at the patent level to map the knowledge space over time. I extend this literature using text to describe patent level innovations to the team level.

Arts, Hou, and Gomez 2021 developed the literature beyond using citations his-

tories by emphasising the use of patents texts to identify novel contributions. Kelly et al. 2021 produced a key development in identifying breakthroughs by comparing the similarity of patent texts to that which came before and after. The concept of breakthrough in this paper builds directly on this foundation. The key contribution of this paper is to extend this to the team level, which given that they are the leading producers of innovation, is key to determining future innovation.

Paper Outline The rest of the paper is structured as follows. Section 2 defines the theoretical framework. Section 3 builds and validates an empirical approximation of the space. The full treatment of the method used can be found in the technical appendix. Section 4 presents a set of descriptive statistics and section 5 presents a conceptual framework to derive a set of hypotheses, the empirical strategy and results. Section 6 concludes.

2 Theoretical Framework

Define $\mathcal{K}$ as a set of K knowledge classes.[2] Each class represents a specialised area of understanding. Inventors produce innovations by combining their knowledge on these classes. I model the innovation and writing of a patent as a single, unified process. There is a fixed vocabulary of words which inventors can use, denoted by V. Inventors use different words when describing different knowledge classes. This is captured by the probability distribution β_k for topic k across the vocabulary. β_{kv} captures the probability of using word $v \in V$ when discussing class k.

A 3-dimensional example is given by

$$\mathcal{K} = \{\text{Computing, Transport, Medicine}\}.$$

The words *hospital, doctor* and *syringe* are more likely to be used when describing a medical innovation than one about transport. One patent though may combine multiple classes. For instance, a drone to deliver prescriptions will likely use words correlated with both the medical and transport classes.

Denote $\Delta(\mathcal{K})$ as the knowledge space which is defined as the $(K-1)$ probability

[2]No two knowledge classes are more similar to each other. This is a simplification that can be addressed with more complex models that allow for correlation between knowledge classes. Consult David M Blei and Lafferty 2005 for further details.

simplex over the set $\mathcal{K}$. θ is a point in the simplex, such that it represents a combination of knowledge classes. Let I be the set of all inventors. Each inventor is characterised by their knowledge profile θ_i. This is drawn from the knowledge space $\Delta(\mathcal{K})$ according to a Dirichlet distribution.

$$\theta_i \sim Dir_{\Delta(\mathcal{K})}(\alpha)$$

Where $\alpha \in \mathbb{R}^K$ is the non-symmetric Dirichlet prior such that $\alpha_k \neq \alpha_j > 0$. The support for a Dirichlet distribution is the set of K-dimensional vectors $\boldsymbol{x}$ where each $x_k \in [0, 1]$ and $\sum_{k=1}^{K} x_k = 1$. The value of the Dirichlet distribution is that each element in the support of a Dirichlet distribution can be treated as a K-dimensional discrete probability distribution.[3]

If the average α_k is low then the mass of the Dirichlet distribution lies in the corners of $\Delta(\mathcal{K})$. This means that inventors are more likely to hold knowledge on a few classes as opposed to being spread over many. In other words, inventors are more likely to be specialists than generalists as the average α_k tends to zero.[4] I allow for a non-symmetric Bayesian prior, so that on aggregate, certain knowledge classes will be more common.

A team $\tau \subseteq I$ is a set of m inventors who produce patent p together. When a team τ collaborates, they first choose the share of the workload to be performed by each team member. These shares are not constrained to be uniform across team members and some may contribute more than others.[5] I model this as a random draw where the team chooses a vector ω_p such that $\sum_{i \in \tau} \omega_{ip} = 1$ and $\omega_{ip} \geq 0$. Each ω_p is drawn uniformly at random. This can be modelled as a draw from another Dirichlet from the set of all possible workload divisions for m team members, denoted as Δ^{m-1}

$$\omega_p \sim Dir_{\Delta^{m-1}}(\mathbf{1}).$$

The team then produces a patent according to the following stochastic process.

[3] In fact the Dirichlet is the conjugate prior for the multinomial distribution, a feature that is utilised in defining the estimation method.

[4] This matches the literature by modelling inventors as more likely to be specialists than generalists.

[5] Inventors are often modelled as agents with a high level of autonomy over project choice and team participation (Akcigit et al. 2018) and allowing for these weights to be chosen optimally is an important next step.

The team first draws the number of words in the patent $N_p \sim G(\cdot)$.[6] Then for each word $w_{ip} = 1, \ldots, N_p$ the team chooses an inventor $i \in \tau \sim \omega_p$ and from that inventor's knowledge distribution chooses a class $k \in \mathcal{K} \sim \theta_i$. Given the corresponding knowledge class to word distribution, the inventor chooses a word $v_{ip} \in V \sim \beta_k$. Each word in the patent is paired with a knowledge class, which produces a patent knowledge class distribution. Since the number of words in a patent is large, in expectation we can define the expected patent knowledge distribution. I denote the expected patent distribution as θ_p^e to simplify notation throughout the paper.

$$\mathbb{E}[\theta_p|\tau, \omega_p] = \sum_{i \in \tau} \omega_{ip}\theta_i = \theta_p^e \tag{1}$$

Therefore, synonymously to inventors, a patent can either be on a very specific topic, or a combination of many. Importantly, inventors, teams and patents now belong to one consistent space. This enables the counting of how much and which type of innovation exists in each local knowledge field.

The knowledge contained in the patent is a function of the inventors who produced it. However, given the stochastic process, the final patent distribution will not equal its expectation: $\theta_p \neq \theta_p^e$. Though it will likely be very close, since the probability that a given team τ produces a patent distribution θ_p is decreasing in

$$d(\theta_p^e, \theta_p) = ||\theta_p^e - \theta_p||^2 \tag{2}$$

The team first assigns roles within the team, which given the knowledge profile of each team member defines the expected outcome of their collaboration. The stochastic process by which the team generates the innovation is consistent with the idea of them pursuing a method of trial and error, in which each inventor tries many ideas and the probability of success is equal to their contribution weight.

Given the previous example of $K = 3$, the knowledge space is a 2-dimensional equilateral triangle and can be represented as in Figure 1. Each of the corners represent perfectly specialised profiles. An inventor or patent may split their knowledge over two of the classes, and hold no knowledge on the third, as in point 4. Point 5 represents the centroid of the simplex, and is a perfect generalist, sharing their

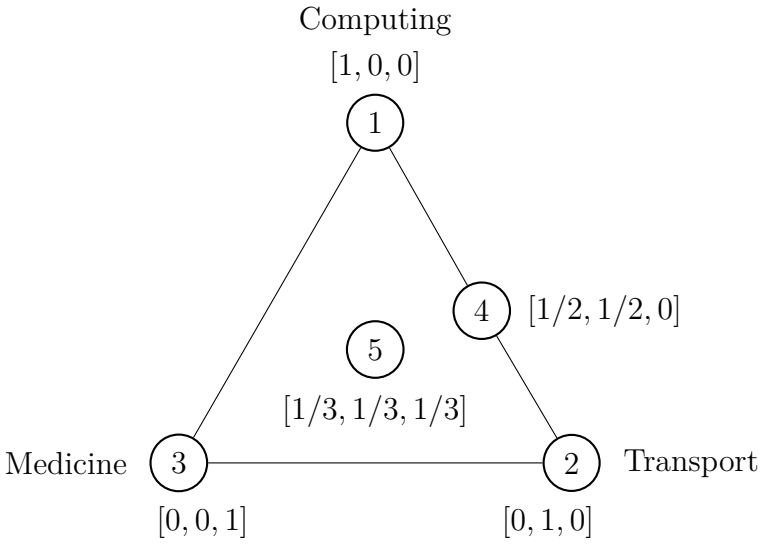
[6]This distribution G is irrelevant for the model. An appropriate approximation can be learnt from the observed set of patent lengths. Potentially this could be interesting over time since patents have become significantly longer throughout the period studied.

knowledge equally over all classes.

If inventors 1 and 2 were to collaborate and contribute equally such that $\omega_{11} = \omega_{21} = 1/2$ then in expectation they will produce θ_p^e at point 4 in Figure 1. Then given the random innovation process, all patents along the line between points 1 and 2 are feasible outcomes, however decreasingly likely as the distance from point 4 increases.

Figure 1

THE KNOWLEDGE SPACE



Notes: An example 2 dimensional knowledge space over 3 knowledge classes. In the full model I use K = 50 classes.

Within this space I define a local knowledge field for both teams and patents. I define a local knowledge field for each patent as a closed ball of radius r centred at point θ given by

$$B(\theta, r) = \{\theta' \in \Delta(\mathcal{K}) \mid \|\theta' - \theta\| \leq r\} \tag{3}$$

This field is fixed over time, however the number of other realised patents belonging to the local knowledge field can vary over time.

I define $S(\tau)$ as the team span: the set of all linear combinations of each team member's knowledge distribution. Given the assumption that the weights ω_p are drawn from a uniform distribution, the team is equally likely to draw any patent in

this set as their expected output, such that $\theta_p^e \in S(\tau)$. Formally I define the team span as the convex hull across team member distributions.

$$S(\tau) = \left\{ \sum_{i \in \tau} \omega_{ip} \theta_i : \sum_{i \in \tau} \omega_{ip} = 1, \omega_{ip} \geq 0 \right\} \tag{4}$$

To define the local knowledge field for a team consider the Minkowski sum of $S(\tau)$ and $B(\theta, r)$. The resulting set is analogous to the local knowledge field at the patent level. In fact the local knowledge field for a team of one is defined identically. This expands the team span into the full K-dimensions of the knowledge space.

$$\tilde{S}(\tau) = S(\tau) \oplus B(\theta, r) = \{x + y \mid x \in S(\tau), y \in B(x, r)\} \tag{5}$$

Continuing with the example outlined previously, Figure 2 demonstrates how inventors, teams and patents lie in one consistent space. Panel (A) shows an example of a patent's local knowledge field. The plot is fixed at the year patent p (shown in black) was published and there were five examples of prior work in that local knowledge field. Panel (B) shows an example team of three members, the blue shaded area represents their span $S(\tau)$. Each inventor lies in one of the vertices of the blue triangle. The red perimeter defines their local knowledge field $\tilde{S}(\tau)$.

When a team τ draws contribution shares ω_p to define their expected patent knowledge distribution θ_p^e within local knowledge field $B(\theta_p^e, r)$. A local knowledge field and time define a breakthrough score (b_p) and innovation direction (z_p) tuple

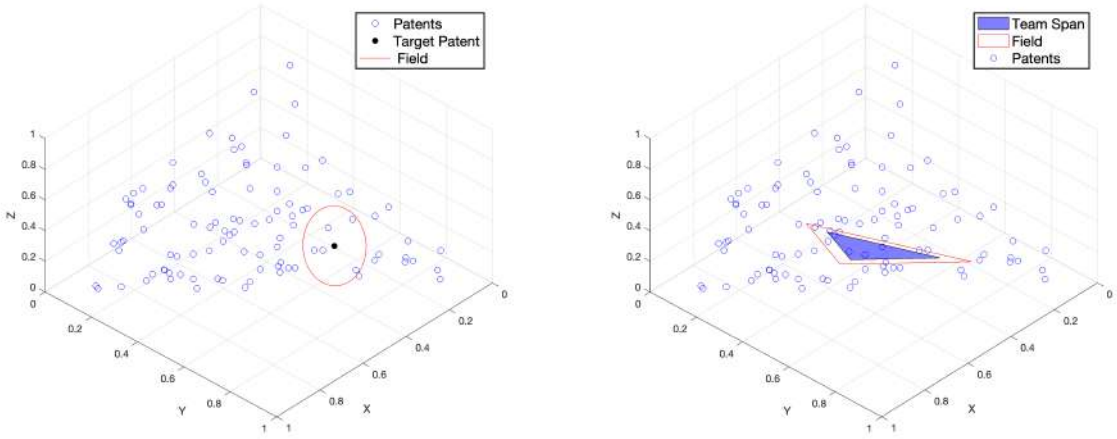
$$(b_p, z_p) \mid \theta_p^e, t.$$

The breakthrough score measures the scientific impact of that innovation. Did it spark a new and successful research field? The direction of an innovation measures the target use of the patent. Does that innovation achieve a certain goal, for example to mitigate climate change, reduce cancer risks or automate production?

The idea being that the impact of an idea is time dependent. The most straightforward example is that there is a significant gain in being the first to invent a new object. If you are working on an artificial intelligence innovations, the same idea has a different value today than it would have had fifty years ago, when many AI models were first theorised. In terms of being a breakthrough, there are now plenty of AI patents which have come before. But the direction—the ability of this combination

Figure 2
LOCAL KNOWLEDGE FIELDS

A) PATENTS B) TEAMS



Notes: The example patents and inventors are generated from a Dirichlet distribution with $\boldsymbol{\alpha} = [2, 1.5, 1]$, which leads to the distribution across the knowledge space being weighted towards the bottom-left corner.

to meet a specific objective—depends on whether similar innovations have previously achieved that goal. If past efforts with similar knowledge combinations have achieved certain outcomes, similar innovations may continue along that path, shaping the future of innovation in that area. Timing plays a critical role, as the same combination might be more or less effective depending on the state of knowledge and technological demand at the time. To complete the prior example, inventors have a wealth of prior AI knowledge to use when automating production today when compared to the past.

Both b_p and z_p are modelled as latent variables, such that for both $y_p \in \{b_p, z_p\}$

$$y_p(\theta_p^e, t) = \begin{cases} 1 & \text{if } f_y(\theta_p^e, t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Allowing for an abuse of notation, f_y is a general function that maps a team's location in knowledge space to the real line. This function can be mapped into the probability that a given patent achieves that outcome. Given the definition of a team span in equation 4, we can define the expected value for each outcome as the following.

$$\mathbb{E}[y_p|\tau, t] = \frac{1}{\text{vol}(S(\tau))} \int_{S(\tau)} y_p(\theta_p^e, t)\, d\theta_p^e \tag{7}$$

In other words, what proportion of all the teams potential ideas achieve outcome y?

Define the set P_t as the set of all patents published across the global knowledge space up to and including period t. Define the following count for the number of these patents which belong to the local knowledge field of a focal patent.[7]

$$n_{pt} = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in B(\theta_p^e, r)\Big) \tag{8}$$

$\mathbb{1}$ denotes an indicator function which is equal to one when the condition in parentheses is met. It is a natural assumption that for a patent to be a breakthrough, it must be one of the first movers. Formally this is given by $\partial f_b / \partial n_{pt} < 0$ and can be tested in the data. Using the expected breakthrough measure for a patent given in equation 7 we can draw the following hypothesis.

Hypothesis 1. *If $\partial f_b / \partial n_{pt} < 0$, a team increases the probability their next patent sparks a breakthrough by reducing the density of patents within their local knowledge field.*

I ask first, how does a team's innovation output develop over time without changing their composition? By fixing the team's position in knowledge space, the volume of their team span remains constant. As teams draw ideas θ_p^e uniformly at random, by increasing the number of other patents within their span, the function f_b decreases over some partition of their potential patents, thus decreasing the expected value.

Part of this result is driven by the team's own innovation. Teams develop on their own patents, potentially transforming their first patent into an established field. But this result also speaks to the literature on the burden of knowledge within the study of endogenous growth (B. Jones 2009). As your local knowledge field populates with patents, it becomes harder to produce a truly innovative idea (Bloom et al. 2020).

In that case, how can a team produce a new breakthrough idea? By altering its composition. This can be achieved either by removing a team member to shift into a sparser region or by adding inventors from less dense areas of the knowledge space. The logic remains the same. Altering the team to collaborate in a sparser, potentially

[7] A detailed explanation of how I count these objects empirically is provided in Appendix B.

smaller area of the knowledge space increases the probability of producing a radical patent.

The rationale for why is as follows. For a team to produce a truly innovative idea, the existence of prior work in the same field is a barrier. Locally to the patent, this is intuitive since it is now not the first to market. At the team level however this result is more subtle. By reducing the density of prior work within their local knowledge field, the team will draw ideas from less populated areas of the knowledge space. The idea being that by reducing the presence of prior work, the team is freed from established paradigms, and are capable of producing a breakthrough idea.

I now move from creating new research fields to the direction of innovation. Define the following count, again over the set P_t, this time including a check for the direction of each patent within a target patent's local knowledge field.

$$n_{pt}(z) = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in B(\theta_p^e, r) \cap z_q = 1\Big) \tag{9}$$

To extend this logic to the direction of innovation I apply the argument behind the ladder of development Grossman and Helpman 1991 from the endogenous growth theory literature. $\partial f_z / \partial n_{pt}(z) > 0$ represents innovation costs decreasing. It becomes less costly for a team to produce a patent of a given type since they develop on prior work. I therefore posit the following hypothesis.

Hypothesis 2. *If $\partial f_z / \partial n_{pt}(z) > 0$, a team increases the probability of producing a patent of type z if they increase the density of patents of type z within their local knowledge field.*

As the field moves up the ladder of development, each step is marginally less costly and teams produce more patents targeting that direction. This speaks to an important literature on pivoting. The arrival of the first patent to their field targetting direction z may cause that team to pivot to also producing patents of that type. Its arrival reduces the pivot penalty by providing a link between the two fields (Hill et al. 2022). Analogous to the hypothesis on producing a breakthrough, a team can proactively change the direction of its patents by altering its composition. By bringing in new members who increase their ability to focus on direction z, this reduces the cost of this type of innovation. This then increases the likelihood of them patenting towards that purpose.

Each result demonstrates how the quantity of prior work determines team output. The effect of prior work depends on the outcome you are measuring. When a team is concerned with whether their innovation achieves a given purpose then prior work reduces the cost of innovating. If you design a particular household object, it is complex to first transfer the knowledge from computer science into your field to produce a smart version. Following on from prior work reduces your costs of combining those knowledge fields, and thus increases the probability of you doing so. When looking to produce a new field however, the existence of prior work is a barrier, both mechanically by not being the first to develop a given product, but also by defining paradigms that guide all future work in that local area.

3 Estimating the Knowledge Space

I first outline the data and sample over which the model is approximated. I then introduce the Bayesian model of Natural Language Processing used to approximate the knowledge space. This allows me to count the quantity and type of prior work within a team's local knowledge field, and combine this with data to test both hypotheses.

3.1 Data and Sample

I estimate the knowledge space on US patent data from *patentsview*, the online data base for the United States Patent and Trademark Office (USPTO). I track teams dynamically and define a panel of team, patent observations. I restrict the sample to teams who applied for their first patent after 1990, and their last prior to 2011. I then stack over various team types which each play a role in the empirical analysis. The first are those teams which are treated by the premature death of a co-inventor. The premature death of an inventor is determined using the dataset provided by Kaltenberg, Adam B Jaffe, and M. Lachman 2021. I define a premature death using the following logic. I take one unique death date per inventor, and classify premature as an inventor who dies within three years of patenting with the team. This defines a treated inventor, and treated team. I then search for teams which return to patent within up to five years in two cases: they return minus the deceased inventor, or having replaced that inventor with one other. Teams which return with two or more new inventors are dropped. Given the delay in producing a patent, less than five

years is relatively fast to turn around a new patent and by controlling for the time I claim that the death was a quasi-natural experiment in changing team composition.

I add to this sample two additional types of teams which act as controls. The first are pure controls: a team which never adds or removes a member. This group of teams never never appear again either without one or more members, or having added one or more new ones. The second are those that first patent with n_τ inventors, then that after that team publishes their final patent the same inventors return, with one additional member, again within up to five years. The set of controls provide a baseline comparison for whether teams change their output dynamically, and the second an endogenous team composition change that allows me to study adding new members. In total I find 353 teams treated by a premature death who return without the deceased inventor, 2200 treated teams that replace that inventor with one other. Then to find the controls from a random sample of 300,000 teams I find 6400 pure control teams and 980 teams which add one new member.

This is the sample used to train the LDA. I define the knowledge space over these teams and patents, but then to measure on what fields do patents build I populate this space with a random draw from the universe of USPTO patents. I extract just over 2.2 million USPTO patents, approximately one third of the universe of USPTO patents grants over the period studied.[8] I populate the knowledge space with this random sample by treating each patent as if it were a new author, who patented one solo paper. Then taking the estimated knowledge class to word distributions I fit each patent into the estimated knowledge space.

I combine additional data for the two patent outcomes, whether they are a breakthrough or achieve a certain direction. Kelly et al. 2021 classify the universe of USPTO patents from 1976-2014 as whether they are a breakthrough, or not. I measure three innovation directions exogenously. They are three binary indicators for whether a given patent achieves that purpose, or not. The first is whether that patent is a labour saving technology (Mann and Püttmann 2023). Secondly does that patent mitigate climate change which is measured as whether that patent is awarded the YO2 patent class (PatentsView 2024 and finally does that patent target improving cancer diagnosis or treatment (USPTO 2024).

[8]This is a rough calculation. To determine the denominator in this calculation I use the fact that there were 6,901,791 patent's granted between 1976 and 2020

3.2 Latent Dirichlet Allocation

Patent texts are increasingly used to describe the knowledge content of innovation, and the innovation literature has begun to borrow and develop models from the computer science literature in order to answer new questions on science and technology. Patent number US9939179 begins their detailed description with the following:

> *However, one of ordinary skill in the art will recognize that the invention is not necessarily limited to refrigeration systems. Embodiments of the invention may also find use in other systems where multiple compressors are used to supply a flow of compressed gas.*

This quote demonstrates that the patent texts are informative on the knowledge content beyond a simple title or CPC classification. In order to extract this information into a empirically feasible dimension I use a model of Latent Dirichlet Allocation (LDA). LDA models were first developed by David M. Blei, Ng, and Jordan 2003 and have become a popular method of NLP. Consider this a brief and intuitive overview of how an LDA infers a set of parameters which approximate the knowledge space. For a full description consult the accompanying technical appendix in C.

The model is built upon the paradigm of observing the set of patent texts, and proposing a hierarchical Bayesian model to infer a set of latent parameters which govern how that set of texts was produced. The model identifies many parameters jointly: The inventor and patent knowledge class distributions and each inventors' contribution weight to each patent.

I build on the *gensim* python package (Mortensen 2017) which trains the unsupervised ML model by implementing a method of Variational Bayes. The objective is to infer from patenting histories which team member was most likely to have contributed each word and with which knowledge class. In doing so, infer the inventor knowledge distributions and their contribution shares to patents. An inventor with a long history of producing transport patents will be more likely to have contributed the words vehicle, destination and route. If a given patent includes many words highly correlated with the transport class, the model will give a larger contribution share to that inventor.

Identification in a Bayesian context is not the same as in frequentist regression models, though there are similarities. The model may converge to a solution and estimate parameters which are not well-identified in the regression context. If two

inventors work together and produce many patents, but only ever working as a pair, it is impossible to disentangle who did what on those patents. In this case the model defaults to an equal probability for each team member across the knowledge classes contained within the patent. This is conceptually equivalent to assigning CPC classes evenly across all team members, therefore in this case the method presented here defaults to the standard method in the literature (Adam B Jaffe 1986). In addition, a topic model makes use of all documents fed into the model to identify the knowledge classes distributions, therefore even if the inventor level parameters are not well-identified, their patents still contribute to estimating other model parameters.

Table 1 provides the hyper-parameters which govern the estimation process.

Table 1
LDA PARAMETERS

K	η	Iterations	Passes	γ
50	1/K	350	100	0.001

Notes: The model has been run various times changing these parameters, and the results are qualitatively similar. Both η and γ are set to the *gensim* default values.

These are the parameters used in estimating the ATM-LDA. η is the prior for the knowledge class to word distribution and is assumed to be symmetric. The number of passes defines the number of times that the model sees the entire dataset, where the number of iterations defines the number of times the model iterates within the EM stage over each document. The model is trained using the online method where documents are loaded in batches of 2000. The choice of $\eta = 1/K = 0.02$ is the gensim default option but also in line with the literature as both Hansen, McMahon, and Prat 2018; Griffiths and Steyvers 2004 set $\eta = 0.025$. Prior to estimating, I preprocess the text in order to improve the model inference, by stemming and removing stopwords Sarica and Luo 2020.

I estimate the Bayesian parameter flexibly instead of defining a fixed prior. This allows for variation in the importance of a knowledge class on aggregate, which reflects a more natural state of the world. The following results are robust to changing the model parameters.[9] Figure S2 plots the log-likelihood and perplexity at each pass over the data which shows that the model converges after approximately 100 passes.

[9]The model has been run with K = 20, 30 and 40 as well as α and η chosen optimally and for a range of iterations, 100, 200, 500.

The model maximises over the variational parameters to minimise the lower bound on the data, in this sense it converges to an approximate solution.

The perplexity measure is the standard measure used within the topic modelling literature to evaluate the quality of topics estimated. The perplexity score measures how well the model predicts the words in the documents based on the learned topic distributions. In other words, how well the model captures the underlying structure of a set of documents. A lower perplexity score indicates that the model has a better ability to generalise to unseen data, and convergence indicates that the LDA has effectively learned the topic structure of the patents.

3.3 Clustering Knowledge Space

The knowledge space is a high dimensional object, which describes the knowledge contained in patents. These patents are objects which are used in a range of industries and markets. In order to control for a range of unobserved heterogeneity I cluster the estimated knowledge space into a set of N clusters, by splitting the knowledge space using K-means clustering. The idea here is that the model is blind to a number of patent characteristics which potentially correlate broadly with the type of knowledge it contains. If there is a correlation between certain classes, or combinations of classes, and high value industries, any regression model studying movements in this space may be biased.

For example, a number of the knowledge classes use words related to information and communication technologies. These industries have seen a number of periods of explosive growth, but also significant changes in regulation and economic outlook. In Figure S3 I provide an example output for the same three dimensional space shown in 2 where the clusters form four broad groups of patents. If patents that are heavily computational patents have on average more citations or a larger market value due to market shocks or time-invariant consumer preferences, this set of cluster controls will account for that.

3.4 Empirical Strategy

I present a set of regression models to test both hypotheses derived in section 2. For the following analysis I use the term *breakthrough model* to refer to modelling the probability that a patent is a breakthrough. The second version models the probabil-

ity that a patent targets one of three directions: automating production, mitigating climate change or reducing cancer risks. This is referred to as the *innovation model*. To analyse the direction of innovation, I stack the three innovation directions into a single regression model to ensure the results remain technologically neutral.

To tackle the research question on how teams build on prior work I first start at the patent level. I test the relationship between the quantity of prior work on which a patent develops and the two innovation outcomes. This is an important step since the hypotheses derived in section 2 make assumptions on the shape of this relationship. Here the dependent variable varies at the patent level, where each patent maps into one team τ and application year t. For both models I use a variation of the regression model specified in equation 10. The regression is run as a logit to model the probability of each outcome $y_{t\tau(p)} \in \{b_p, z_p\}$

$$Pr\Big(y_{t\tau(p)} = 1 \mid X'_{t\tau(p)}\boldsymbol{\psi}\Big) = \frac{\exp\Big(X'_{t\tau(p)}\boldsymbol{\psi}\Big)}{1 + \exp\Big(X'_{t\tau(p)}\boldsymbol{\psi}\Big)}$$

where

$$X'_{t\tau(p)}\boldsymbol{\psi} = \beta_0 + \beta_1 x_{t(p)} + \beta_2 d_p + \beta_3 \cdot t + X'_{\tau t}\mu + \delta_c + \delta_z \tag{10}$$

For $x_{pt} \in \{n_{pt}, n_{pt}(z)\}$, either the count of all patents belonging to the local knowledge field of patent p, as defined in equation 8, or over patents of direction z as defined in equation 9. The main parameter of interest is β_1. Where to match the assumptions made in each hypothesis, I require the following sign for β_1. For the breakthrough model I assume that $f_n/\partial n_{pt} < 0$, which corresponds to $\beta_1 < 0$. Vice versa for the direction model I assume that $f_z/\partial n_{pt}(z) > 0$, which requires $\beta_1 > 0$.

The model controls for the randomness in innovation by including the distance between the realised patent distribution and the expected value in $d_p = d(\theta^e_p, \theta_p)$ as defined in equation 2. I include a set of team controls in $X_{\tau t}$, which include a measure of volume for the team span. I then include a set of fixed effects for knowledge cluster (c) as defined in section 3.3 and, when required, direction (z). Introducing a time trend through β_3 achieves multiple purposes. It controls for the fact that breakthroughs are right-coded in time: patents published recently have not yet had chance to be realised as breakthroughs. Also for the fact that patents increasingly tend to achieve all three directions more over time.

I then test hypothesis 1 and 2 for two cases. In the first team composition is held constant. For the second team composition is changed following the premature death of a team member using the data provided by Kaltenberg, Adam B Jaffe, and M. Lachman 2021. I define a premature inventor death as someone who dies within three years of having applied for a patent. Treated teams are those that return to patent within 5 years of the inventors death, either without replacing the inventor (strong exogeneity) or replacing them with one new inventor (weak exogeneity). Both provide a plausibly, though varying in strength, exogenous change in team composition. I don't condition on age at death, instead focusing on them being recently active. Using the premature death of scientists has become a well established source of exogeneity in collaboration (Azoulay, Fons-Rosen, and Graff Zivin 2019; Azoulay, Graff Zivin, and J. Wang 2010). I drawn upon the arguments presented by this important literature on the validity of this method.

All variables are defined analogously as in equation 10, except $n_{\tau t}$ and $n_{\tau t}(z)$ now count the number of patents within the team's local knowledge field. To be clear, I define $n_{\tau t}$ as the sum over all patents published in the knowledge space prior to t (P_t) and count those that belong to the local knowledge field of team τ ($\tilde{S}(\tau)$).

Equation (11) measures changes in team output, holding team composition fixed.

$$X'_{s\tau(p)}\boldsymbol{\psi} = \alpha_\tau + \delta_{sz} + \beta_1 n_{\tau t} + \beta_2 d_p + X'_{\tau t}\mu + \beta_3 \cdot t + \delta_c \tag{11}$$

This is run as a TWFE regression, conditional on the team identifier τ. Each period in this model refers to a new patent, not the year it is produced. Such that δ_{pz} corresponds to the patent order (1,2,3 etc.) in each direction. I include the time trend to control for the systematic changes over time, as previously discussed. When running the breakthrough model there is no direction dimension and $\delta_{sz} = \delta_s$. The idea here is to test the response of team output to changes in their composition, controlling for the year in which they patent. The patent order is an important control, especially when considering breakthrough innovations since teams build on their own prior work. Again in this model, to support hypothesis 1 and 2, I predict that $\beta_1 > 0$ for the direction model and $\beta_1 < 0$ for the breakthrough model.

The headline result is how team innovation outcomes change after moving into a new area of the knowledge space and therefore building on a different set of prior work. I utilise two types of changes to identify the effect of shifting the location of a team. Both follow the premature death of a team member, in which the team

returns to patent within 5 years, denoted as the new team τ'. Either τ' consists of the original team minus the deceased inventor ($\tau' = \tau/\{i\}$), or they replace i with one other inventor j ($\tau' = \tau/\{i\} \cup \{j\}$).

I define the measure $D_{\tau' t} = n_{\tau t} - n_{\tau' t}$ to measure the change in the quantity of prior work on which the team is building, following their shift in the knowledge space. $D_{\tau' t}(z)$ is analogous but only counts those patents targeting direction z. I control in the regression for the first teams count $n_{\tau t}$, such that β_1 captures the effect of removing existing patents from the team span, conditional on the prior quantity. The hypothesis requires that the density of patents changes within the team span. Therefore I introduce a control for the volume denoted v_τ. I approximate the volume of a team span using the following

$$v_\tau = \sqrt{m} \times (\alpha \cdot D_{\max} + (1 - \alpha) \cdot D_{\text{mean}})) = \text{vol}\Big(S(\tau)\Big)$$

Where m denotes team size and $D_{\max}$ is the maximum distance between any two team member distributions, and D_{mean} is the average across all pairwise combinations of team members.

$$X'_{s\tau(p)}\boldsymbol{\psi} = \alpha_{\tau'} + \delta_{sz} + \beta_1 D_{\tau' t} \cdot \mathbb{1}_{post} + \beta_2 v_{\tau'} + \beta_3 n_{\tau t} + \beta_4 d_p + \beta_5 \cdot t + X'_{\tau t}\boldsymbol{\mu} + \delta_c \quad (12)$$

β_1 captures how team output changes in response to reducing the quantity of prior work on which a team develops, holding the volume constant across teams. This therefore reduces the density. $D_{\tau' t}$ is defined such that for both the breakthrough and direction hypotheses, the coefficient β_1 is predicted to be positive.

4 Describing the Knowledge Space

In this section I present a set of new descriptive statistics which are feasible in the knowledge space and provide important insights into team innovation. I also use this as a chance to validate the space by comparing results from the estimated model to data taken from the literature.

4.1 Breakthrough Patents

In Section 2 I define a local knowledge field spatially as the area around the knowledge distribution of a target patent, defined by a radius r. For any patent, define the time t as the year in which that patent was applied for. I propose the following breakthrough measure at the patent level, which is an adjusted percentage change to allow for zero patents either before, after or both.

$$b_p = \left(\frac{\text{post-count}_p}{1 + \text{prior-count}_p + \text{post-count}_p} \right) \times 100 \tag{13}$$

Prior-count$_p$ counts the number of patents which already existed in the space $B(\theta_p, r)$ prior to this patent's publication in t, and post-count$_p$ the number which came after. Holding prior-count$_p$ constant, the breakthrough score of a given patent p is increasing in the number of patents which came afterwards. It increases non-linearly, with decreasing returns, such that early entrants contribute more than late comers. Figure S1 gives an example that also demonstrates that the curve b_p with respect to post-count$_p$ flattens as the prior-count$_p$ increases.

The measure presented in equation 13 is the raw breakthrough measure, however as made clear in Hall, Trajtenberg, and Adam B. Jaffe 2001, when working with patent outcomes it is important to control for the fact that they are right-coded in time. Patents produced recently have not had enough time to be revealed as breakthroughs, since the patents that build on them have not yet arrived. Therefore, unless where stated, instead of the raw measure I use the residuals from a regression of the breakthrough measure b_p on a set of patent application year dummies.

This spatial definition of a local knowledge field is static over time, however, it leads to a natural concept for a breakthrough patent. I compare the number of patents which existed in the local field prior to p being produced, to those that came after. Patents produced in areas with few pre-existing works are novel, but only those which post-publication see a significant increase in the number of patents belonging to their local knowledge field are breakthroughs. This is similar in concept to the breakthrough measure proposed by Kelly et al. 2021, however uses a spatial dimension that is easier to track over time.

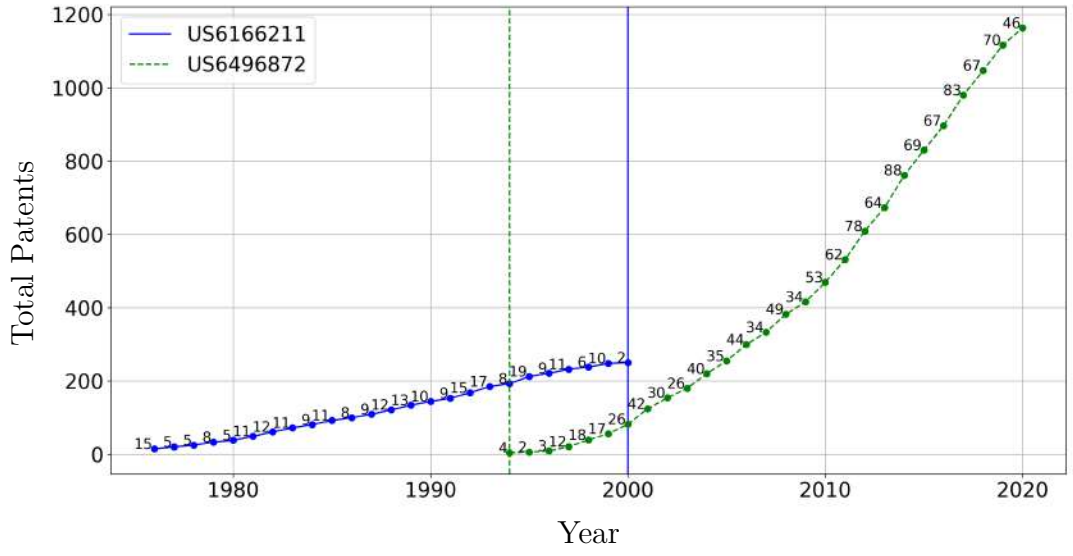
The concept is similar to the previous literature which defines a pre- and post-patent period to capture patent importance. One key contribution of this paper is to extend this to the teams which produce these patents. Figure (3) provides two

examples as the patents corresponding to the minimum and maximum b_p values. The vertical line correpsondes to the year in which each patent was produced. The Y-axis plots the total number of patents within each patent's local knowledge field. Patent *US6496872* titled *Computer system for automatically instantiating tasks designated by a user* which scores very highly given it was one the four first movers in an area with no pre-existing patents, and many patents joined its field after publication. Whereas for patent *US6166211* titled *Manure-spreader* they develop on an area showing slow growth, and no futher patents came after them.

I lean on the existing literature using NLP to produce similarity measures between patents by calculating the backward and forward similarity of patents to validate my method. I classify breakthrough patents as those that land in the top decile, where I use the residuals from a regression of b_p on a set of application year dummies to clean out the issue of right-coding of patent data over time.

Figure 3

Evolution of local knowledge fields



Notes: Examples of low and high breakthrough patents. Patent US6496872 is a breakthrough patent, where patent US6166211 is not. This is the raw data, and doesn't remove year effects by normalising over time. The vertical line identifies the publication year for each patent.

Table 2 provides a set of validation statistics to demonstrate the empirical power of the framework. This paper develops on the work in Kelly et al. 2021 and using their data I find the correlation between their binary breakthrough classification and the one produced in this paper. I find a positive correlation of 0.221. I extend this and

show that their continuous breakthrough score is negatively correlated with the prior-count of patents belonging to that local knowledge field, but positively correlated with the post-count.

In addition, using the Arts, Hou, and Gomez 2021 data I first show that patents which I classify as breakthrough patents contribute 5.47% more new words which then go on to be re-used by future patents. This is a straightforward example of creating a new research field. They also introduce significantly more new combinations of existing words, 26.1% new word pairs, and 26% new-three word tuples.

Table 3
VALIDATION OF BREAKTHROUGH PATENTS

Kelly et al. (2021)	Correlation between breakthrough measures	0.234^{***}
	Corr. between pre-count$_p$ and breakthrough score	-0.121^{***}
	Corr. between post-count$_p$ and breakthrough score	0.226^{***}
Arts et al. (2021)	$\%\Delta$ new re-used words in breakthrough patents	8.67^{***}
	$\%\Delta$ new re-used bi-grams in breakthrough patents	47.6^{***}
	$\%\Delta$ new re-used tri-grams in breakthrough patents	44.1^{***}
Citations	$\%\Delta$ forward citations for + $\Delta 1\%$ in post count$_p$	$2.07\%^{***}$
	$\%\Delta$ backward citations for + $\Delta 1\%$ in prior count$_p$	$1.09\%^{***}$
	$\Delta\%$ forward citations for breakthrough patents	$16.2\%^{***}$

Notes: Validation statistics using UPSTO citation data and existing patent novelty literature. The correlation between the Arts, Hou, and Gomez 2021 and Kelly et al. 2021 is 0.28 for backward similarity and 0.29 for forward similarity. The average number of new words, bi-grams and tri-grams used is 1.53, 5.85 and 8.08 respectively. The first panel displays the pairwise correlation coefficient. The second and third panels present log-log regression coefficients from a model which controls for application year and cluster dummies.

Finally, I find that for each additional 1% of patents to enter the local knowledge field of a patent after its publication, the target patent receives 2.49% more citations. This elastic response points to the existence of knowledge spillovers between local patent sub-fields. This logic also holds for backwards citations where for each additional 1% of patents already present in a local knowledge field when a patent is produced, the target patent makes 1.59% more backward citations.

4.2 Innovation Direction

The words contained in a patent describe its design and use. The method reduces the dimension from over 250,000 words to infer a distribution for each knowledge class across the set of unique words. The logic here is that certain knowledge fields use specific words, jargon, more than others when describing objects or problems from their field. For example, someone describing a medical patent is more likely to use the words blood, cells and syringe than someone talking about vehicles, who is more likely to use car, wheel and door.

The model uses the knowledge classes as a dimension reduction technique since a distribution for all inventors across all words is harder to manage both conceptually and computationally. For the following example I set $K = 50$ prior to estimating the model. In appendix section E I show the word clouds for each of the 50 knowledge classes. The words presented are stemmed as part of the text cleaning process, e.g. the word *imag* represents image, images and imaging. The model does not attach labels to the knowledge classes, though they can be approximated using GPT technologies which analyse the word weights.

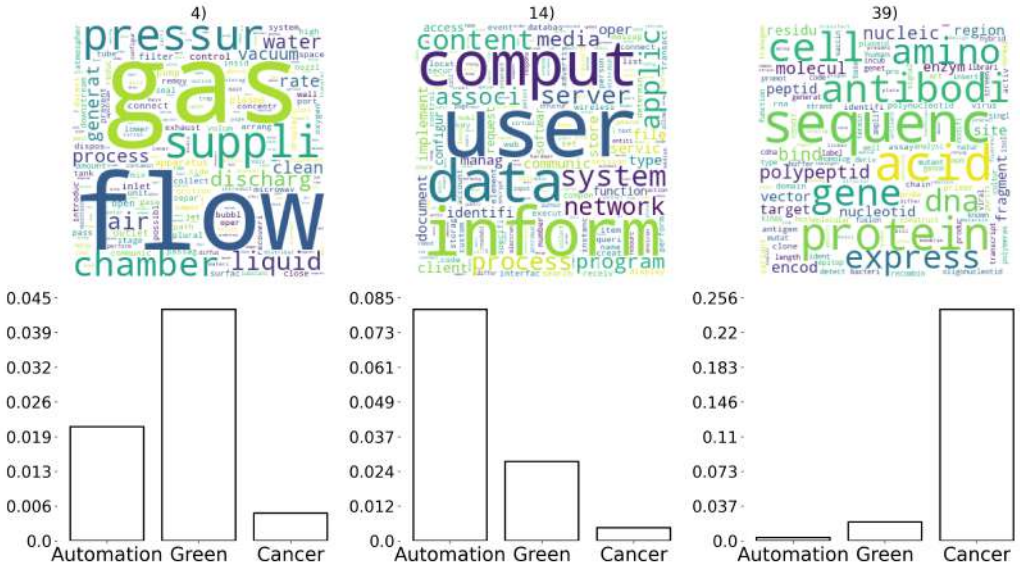
Innovation is used to solve some of societies greatest challenges. For this paper I take three exogenous classifications of whether each patent in the knowledge space is classified as whether that patent achieves that purpose, or not. The directions are the following. Does the patent save labour? Does the patent reduce cancer risks or improve treatment? Does the patent mitigate the negative effects of climate change? These classifications are designed to be industry and technology neutral, in that patents using a broad range of underlying knowledge can target any of these (non-mutually exclusive) directions.

Figure S5 plots the estimated Bayesian prior over the knowledge classes and the 5 words with the largest weight within the distribution for that class. We see variation across classes, which allows for some classes to be over-represented, which will reflect aggregate innovation direction across the time period.

Given that each patent is classified by whether that patent achieves that purpose, I can examine how variation in the words used determined the purpose of the patent. For example, by comparing the most frequent topics across patents that mitigate climate change, target cancer treatment or produce artificial intelligence, we can see how these purposes are achieved. Figure S6 shows the average weight for three knowledge classes (realised) split over three patent types.

Figure 4

WORDCLOUDS AND KNOWLEDGE CLASS DISTRIBUTIONS BY PATENT TYPE



Notes: The bar chart shows the mean weight on a select three of the fifty knowledge classes, averaged across patents of each type. These types are not mutually exclusive. The word cloud is plotted using the estimated knowledge class to word distributions.

Previously the literature has modelled knowledge over time through citations. This Figure demonstrates the power of the model proposed by this paper compared to using citations. By estimating a fixed space, I can plot the evolution of a knowledge field over time. Using citations there is no fixed distance measure and knowledge fields are defined endogenously by which patents cite each other. This also applies to alternative text analysis methods that use the similarity between patent texts to measure local knowledge fields.

4.3 Contribution Weights

This paper is the first to estimate the contribution of each team member to the knowledge contained in a patent. To demonstrate the power of this method, I validate the inventor contribution weights using a prediction model. I propose that if the weights capture information on the true contribution share of each inventor, then the patenting history of inventors who contribute significantly more should be a stronger determinant of the technology classification awarded to a patent.

For each patent in the sample I define the lead and second inventor by ordering

their estimated contribution shares and calculate the percentage difference between them. With a random forest, I predict the CPC classifcation awarded to a patent with two sets of explanatory variables: the five most common CPC classes used by the lead inventor, prior to the target patent, and the corresponding five for the second inventor. When using a random forest you can then calculate the feature importance for each explanatory variable, similar in concept to measuring how each variable contributes to the R^2 of a regression.

I propose that if the gap between the contribution shares of the two inventors is large, then the lead inventor's patenting history will be a significantly stronger predictor of the CPC class awarded to a patent. While if that difference is small (both inventors contributed similarly to the patent), then I predict there to be no significant difference. This correspondes to the total feature importance for the lead inventor's patenting history being significantly larger than that of the second inventor.

Table 2

VALIDATION OF THE CONTRIBUTION WEIGHTS

	$\%\Delta \geq p90$		$\%\Delta \geq p75$		$\%\Delta \leq p25$		$\%\Delta \leq p10$	
T-Test	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Lead	57.126	0.019	56.806	0.015	50.244	0.045	50.030	0.031
Second	42.874	0.019	43.194	0.015	49.755	0.045	49.970	0.031
Difference	14.251	0.027	13.612	0.021	0.488	0.064	0.059	0.043

Notes: T-test to determine differences across lead and second inventor feature importance. Small and large gaps are defined by the percentiles on the percentage difference $\%\Delta$ between the lead and second inventor. After each run of the random forest I calculate the total feature importance for the lead and second inventor patent histories such that the final T-test is calculated over N=50.

Table 3 shows a T-test over 50 runs of a random forest, where each run I draw a new split of the training and testing data set. This is a form of cross-validation that removes the dependency of the outcome on a random initial seed and allows me to estimate a standard error. The null hypothesis for the T-test is that both the lead and second inventor contributed equally.

I find that for teams in which the lead inventor contributes substantially than the second inventor (top 10 or 25%), their patenting history is around 14 percentage points more informative about the CPC classification their joint patent is be awarded. When conditioning on the difference between the first and second inventor being small, this

difference disappears, which points to the contribution weights providing economically important and precise information on who contributed to the knowledge contained.

5 Main Results

I present the main results to test the hypotheses laid out previously.

5.1 Teams that Spark New Research Fields

In these regressions the outcome variable is taken from the Kelly et al. 2021 data.[10] I take their breakthrough measure instead of the one developed in this paper to remove concerns that the effect is mechanically driven by the definition in equation 13.

[10] I use their breakthrough measure calculated on the 90 percentile, based on the previous 5 years. When using the other variations they calculate, results remain similar.

Table 4
PATENT REGRESSION ESTIMATES

I: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
Prior work$_{pt}$	-0.003***	-0.005***	-0.002***	-0.002***
	(-5.64)	(-10.20)	(-6.44)	(-6.42)
N	321140	321140	321140	321140
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time Trend			✓	✓
Team size				✓

II: DIRECTION

	Dependent variable: Pr(Direction)			
Prior work | Direction$_{pt}$	0.022***	0.022***	0.022***	0.022***
	(15.41)	(18.11)	(17.94)	(17.97)
N	1218385	1218385	1218385	1218366
Controls	✓	✓	✓	✓
Direction FE	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time Trend			✓	✓
Team size				✓

Notes: Each column corresponds to a logistic regression of the probability a patent is either a breakthrough (b) or one of three types (z), where all three types are stacked into one regression model. In panel I) the dependent variable is the probability that patent is in the top 90% in the Kelly et al. 2021, for 5 years. In panel II) the dependent variable is composed of three binary indicators for whether that patent achieves each of the three directions: mitigates climate change, reduces cancer risk or automates production. All standard errors are clustered at the knowledge cluster × year level. Controls include $d(\theta_p^e, \theta_p)$. If you use the full count, instead of splitting by direction, then the coefficient is negative.

All variables are defined as in section 3.4. All regression tables show a set of regression models that increase in rigour in each additional column, I interpret all results taken from the final column. From Table 4 panel I) I confirm that at the patent level, the probability a patent is a breakthrough is a decreasing function of the number of pre-existing patents within its local field. Each additional pre-existing local patent is associated with a 0.002% decrease in the odds that patent becomes a breakthrough.

Table 5
WITHIN TEAM REGRESSION ESTIMATES

I: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
Prior work$_{\tau t}$	-0.025***	-0.024***	-0.053***	-0.018
	(-4.50)	(-4.24)	(-4.94)	(-1.63)
N	442	442	442	442
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls		✓	✓	✓
Cluster FE			✓	✓
Time trend				✓

II: DIRECTION

	Dependent variable: Pr(Breakthrough)			
Prior work | Direction$_{\tau t}$	0.056***	0.056***	0.057***	0.057***
	(20.51)	(20.51)	(20.53)	(20.53)
N	6078	6075	6075	6075
Team FE	✓	✓	✓	✓
Period × Direction FE	✓	✓	✓	✓
Controls		✓	✓	✓
Cluster FE			✓	✓
Time trend				✓

Notes: All regressions are fixed effect regression models run with `xtlogit`. The dependent varaible for panel I) is a binary indicator for whether the that patent is a breakthrough, or not. The dependent variable for panel II) is again a stacked binary indicator for whether a patent achieves the given direction: mitigates climate change, reduces cancer risk or saves labour. All regressions include team and patent order fixed effects and standard errors are clustered at this level. Controls include $d(\theta_p^e, \theta_p)$.

In table 5 I first present supporting evidence for hypothesis 1. Holding the team fixed, as more patents arrive, the probability they produce a breakthrough decreases. Each additional patent which arrives to the team local knowledge field is associated with a decrease in the odds that the patent becomes a breakthrough of 0.05%, however importantly this effect disappears when controlling for the time trend.

Table 6
TREATMENT TEAM REGRESSION ESTIMATES

I: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
$D_{\tau' t} \cdot \mathbb{1}$	0.004*	0.006**	0.004*	0.002
	(2.27)	(3.06)	(2.19)	(1.42)
Prior work$_{\tau t}$	-0.017***	-0.031***	-0.013*	-0.012*
	(-4.21)	(-5.17)	(-2.37)	(-2.12)
Volume$_{\tau}$				-2.312**
				(-2.78)
N	601	601	601	601
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

II: DIRECTION

	Dependent variable: Pr(Direction)			
$D_{\tau' t}$(Direction) $\cdot \mathbb{1}$	-0.009***	-0.009***	-0.008***	-0.009***
	(-4.68)	(-4.66)	(-4.53)	(-4.53)
Prior work | Direction$_{\tau t}$	0.042***	0.043***	0.043***	0.043***
	(22.28)	(22.30)	(22.36)	(22.37)
Volume$_{\tau}$				-0.212
				(-0.93)
N	6666	6666	6666	6666
Team FE	✓	✓	✓	✓
Period × Direction FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

Notes: All regressions are team and patent order fixed effect models and standard errors are clustered at this level. The dependent variable for panel I) is an indicator for whether the patent is a breakthrough using the Kelly et al. 2021 data. The dependent variable for panel II) is a stacked indicator for whether a patent achieves the given direction: mitigates climate change, reduces cancer risk or automates production. Controls include $d(\theta_p^e, \theta_p)$.

Finally, Table 6 reports the coefficient of interest on the response of a team's patents to changing the team composition. Here the results in column 4 are weaker than expected, however all coefficient signs support hypothesis 1. When a team moves into a sparser section of the knowledge space, they increase their probability of producing a breakthrough idea. The coefficient on the team span volume is negative and highly significant. This is of interest since this matches the literature on how small teams are more likely to produce breakthrough ideas Wu, D. Wang, and Evans 2019.

5.2 The Direction of Team Innovation

For these results the outcome variable is the classification of patents according to the three types. Do they save labour, mitigate climate change or reduce the risks of cancer. These three are not mutually exclusive. Recall that this corresponds to a stacked regression, controlling for each direction through a direction fixed effect.

From Table 4 panel II) I confirm that at the patent level, the probability a patent achieves a given direction is an increasing function of the pre-existing number of patents of that type in its local field. Each additional pre-existing and local patent of that type is associated with a 0.02% increase in the odds that patent achieves direction z.

Given this result, in Table 5 I first present supporting evidence for hypothesis 2 without changing team composition, therefore holding the volume of the team span constant. Each additional patent which arrives to the team local knowledge field is associated with an increase in the odds that the patent targets direction z of 0.05%.

Finally, Table 6 reports the coefficient of interest on the response of a team's patents to changing the team composition. For each additional patent lost from the team knowledge field, the probability their next patent targets direction z decreases by 0.008%. While this may seem small, the average number of patents lost when a team doesn't replace the inventor is 24.63, which leads to decrease in the expectation of $Y_{p(\tau t)}$ of 1.14% [11]. When they do replace the inventor, they on average gain 3.06 patents, which essentially closes the loss.

[11] $(24.63 \times 0.008)/\ 0.172 = 1.14$

5.3 Extensions and Robustness

I present an extension that opens the door to a dynamic model, and a discussion of the robustness tests run. The tables for the robustness tests can be found in D.

5.3.1 Learning Effects

Table 7
LEARNING EFFECTS

I: BREAKTHROUGH		II: DIRECTION	
	Pr(Breakthrough)		Pr(Direction)
$D_{\tau' t} \cdot \mathbb{1}$	0.015*	$D_{\tau' t}(z) \cdot \mathbb{1}$	-0.013***
	(2.39)		(-4.55)
$D_{\tau' t} \cdot \mathbb{1} \cdot p_\tau$	-0.006*	$D_{\tau' t}(z) \cdot \mathbb{1} \cdot p_\tau$	0.002*
	(-2.05)		(2.18)
Prior work$_{\tau t}$	-0.018**	Prior work | Direction$_{\tau t}$	0.043***
	(-2.72)		(22.32)
Volume$_\tau$	-2.307**	Volume$_\tau$	-0.180
	(-2.65)		(-0.79)
N	601	N	6666
Team FE	✓	Team FE	✓
Period FE	✓	Period × Direction FE	✓
Controls	✓	Controls	✓
Cluster FE	✓	Cluster FE	✓
Time Trend	✓	Time Trend	✓

Notes: All regressions are fixed effect regression models run with `xtlogit`. The dependent varaible for panel I) is a binary indicator for whether the that patent is a breakthrough, or not. The dependent variable for panel II) is again a stacked binary indicator for whether a patent achieves the given direction: mitigates climate change, reduces cancer risk or saves labour. All regressions include team and patent order fixed effects and standard errors are clustered at this level. Controls include $d(\theta_p^e, \theta_p)$. The treatment $D_{\tau' t}$ counts the difference in quantity of prior work between the local knowledge field of a team before (τ) and after (τ') the premature death of a member. p_τ counts the number of patents the team produced before the prematur death as team τ.

The model presented in this paper does not allow inventor knowledge profiles to update dynamically. An important avenue for future work is to develop a dynamic version of the model, that allows for inventors to learn new knowledge over time. The literature on LDAs has developed a dynamic version of the standard topic model which allows for knowledge classes to update over time (David M Blei and Lafferty 2006). This demonstrates the potential for developing a dynamic author-topic model.

For now, I present a simple test to open the door to future research. I introduce into equation 12 an interaction between the continuous treatment $D_{\tau' t}$ and the number of patents that the first team, including the inventor who dies prematurely, produced together. Denote this patent count as p_{τ}. I propose that teams which lose an inventor, with whom they have a long patenting history, suffer a reduced impact of their premature death. Importantly this is in terms of the knowledge contributed by that team member. Jaravel, Petkova, and Bell 2018 show that research teams suffer a significant loss of accumulated research capital when a team breaks up. I propose however that conditional on the team returning to patent, the loss of an inventor is mitigated if the team has had time to learn from one another. Therefore the loss of patents from their local knowledge field is less severe. Technically, this corresponds in a positive coefficient for the direction model, and a negative one for the breakthrough, mitigating the impact of the premature death.

Table 7 indicates that team members potentially learn from each other as both the coefficient on the three-way interactions match the proposed sign. The coefficient on the probability of producing a breakthrough is negative. This result should be taken cautiously. As one of many examples of the potential for future work, a dynamic version of the model would allow for a much more rigorous treatment.

6 Concluding Remarks

This paper presents a novel framework for modelling knowledge production. The paper builds a mapping of inventors, teams and patents in which to study how teams innovate. I approximate this mapping by developing a model of Bayesian Natural Language Processing. As the first to integrate inventors and patents into one consistent space, the paper re-conceptualises how knowledge is produced by recombining existing knowledge and standing on the shoulders of giants. The paper contributes a greater understanding of the key latent variables behind knowledge production and al-

lows me to tackle a set of important hypotheses on how team composition determines innovation outcomes.

The position of a team within knowledge space is a key determinant of their innovation output. Teams which occupy dense areas, in which there exists a wealth of prior knowledge, naturally find it harder to produce a breakthrough idea. However if that prior work targets a specific purpose, they are more likely to target that same direction, given that they have stronger foundation on which to build. By adding and removing members, the team can increase their probability of producing a breakthrough or innovations targeted at certain direction.The framework presented here opens a world of future research. I hope that others are encouraged to utilise this framework to continue deepening our understanding of how and why we produce science and technology.

References

Ahmadpoor, Mohammad and Benjamin F. Jones. "Decoding team and individual impact in science and invention". In: *PNAS* 116.28, pp. 13885–13890.

Akcigit, Ufuk et al. *Dancing with the Stars: Innovation Through Interactions*. Working Paper 24466. National Bureau of Economic Research.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez. "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures". In: *Research Policy* 50.2, p. 104144.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin. "Does Science Advance One Funeral at a Time?" In: *American Economic Review* 109.8, pp. 2889–2920.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. "Superstar Extinction". In: *The Quarterly Journal of Economics* 125.2, pp. 549–589.

Blei, David M and John D Lafferty. "Correlated topic models". In: *NeurIPS*, pp. 147–154.

Blei, David M and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Bloom, Nicholas et al. "Are Ideas Getting Harder to Find?" In: *American Economic Review* 110.4, pp. 1104–44.

Boerma, Job, Aleh Tsyvinski, and Alexander P Zimin. *Sorting with Team Formation*. Working Paper 29290. National Bureau of Economic Research.

Bonhomme, Stéphane. "Teams: Heterogeneity, Sorting, and Complementarity". In: *Unpublished manuscript*.

Devereux, Kevin. *Identifying the value of teamwork: Application to professional tennis*. Working Paper Series 14. University of Waterloo.

Fleming, Lee and Olav Sorenson. "Science as a map in technological search". In: *Strategic Management Journal* 25.8-9, pp. 909–928.

Freund, Lukas. *Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities*. Working Paper. SSRN.

Griffiths, Thomas L. and Mark Steyvers. "Finding Scientific Topics". In: *Proceedings of the National Academy of Sciences* 101.1, pp. 5228–5235.

Grossman, Gene M and Elhanan Helpman. "Quality Ladders in the Theory of Growth". In: *Review of Economic Studies* 58.1, pp. 43–61.

Hall, Bronwyn, Manuel Trajtenberg, and Adam B. Jaffe. *The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools*. Working Paper 3094. Centre for Economic Policy Research.

Hansen, Stephen, Michael McMahon, and Andrea Prat. "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach". In: *Quarterly Journal of Economics* 133.2, pp. 801–870.

Herkenhoff, Kyle et al. "Production and Learning in Teams". In: *Econometrica* 92.2, pp. 467–504.

Hill, Ryan et al. "Adaptability and the Pivot Penalty in Science". In: *Unpublished manuscript*.

Jaffe, Adam B. "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value". In: *American Economic Review* 76.5, pp. 984–1001.

Jaravel, Xavier, Neviana Petkova, and Alex Bell. "Team-Specific Capital and Innovation". In: *The American Economic Review* 108.4-5, pp. 1034–1073.

Jones, Benjamin. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" In: *The Review of Economic Studies* 6 (1), pp. 283–317.

Kahane, Leo, Neil Longley, and Robert Simmons. "The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and Language Diversity in the National Hockey League". In: *The Review of Economics and Statistics* 95.1, pp. 302–314.

Kaltenberg, Mary, Adam Jaffe, and Margie E. Lachman. *Matched inventor ages from patents, based on web scraped sources*. Harvard Dataverse. URL: https://doi.org/10.7910/DVN/YRLSKU.

Kaltenberg, Mary, Adam B Jaffe, and Margie Lachman. *The Age of Invention: Matching Inventor Ages to Patents Based on Web-scraped Sources*. Working Paper 28768. National Bureau of Economic Research.

Kelly, Bryan et al. "Measuring Technological Innovation over the Long Run". In: *American Economic Review: Insights* 3.3, pp. 303–20.

Mann, Katja and Lukas Püttmann. "Benign Effects of Automation: New Evidence from Patent Texts". In: *The Review of Economics and Statistics* 105.3, pp. 562–579.

Mortensen, Olavur. "The Author Topic Model". In: *Unpublished manuscript*.

PatentsView. *USPTO Patent Data for Inventors and Assignees*. United States Patent and Trademark Office (USPTO). URL: %7Bhttps://patentsview.org%7D.

Pearce, Jeremy. "Idea Production and Team Structure". In: *Unpublished manuscript*.

Rosen-Zvi, Michal et al. "The Author-Topic Model for Authors and Documents". In: *CoRR* abs/1207.4169, pp. 487–494.

Sarica, Serhad and Jianxi Luo. "Stopwords in Technical Language Processing". In: *CoRR* abs/2006.02633.

Teodoridis, Florenta, Jino Lu, and Jeffrey L Furman. *Mapping the Knowledge Space: Exploiting Unassisted Machine Learning Tools*. Working Paper 30603. National Bureau of Economic Research.

USPTO. *Cancer Moonshot Patent Data*. URL: https://www.uspto.gov/ip-policy/economic-research/research-datasets/cancer-moonshot-patent-data.

Uzzi, Brian et al. "Atypical Combinations and Scientific Impact". In: *Science* 342, pp. 468–472.

Weidmann, Ben and David J. Deming. "Team Players: How Social Skills Improve Team Performance". In: *Econometrica* 89.6, pp. 2637–2657.

Weitzman, Martin L. "Recombinant Growth". In: *The Quarterly Journal of Economics* 113.2, pp. 331–360.

Wu, Lingfei, Dashun Wang, and James A Evans. "Large teams develop and small teams disrupt science and technologyt". In: *Nature* 566, pp. 378–382.

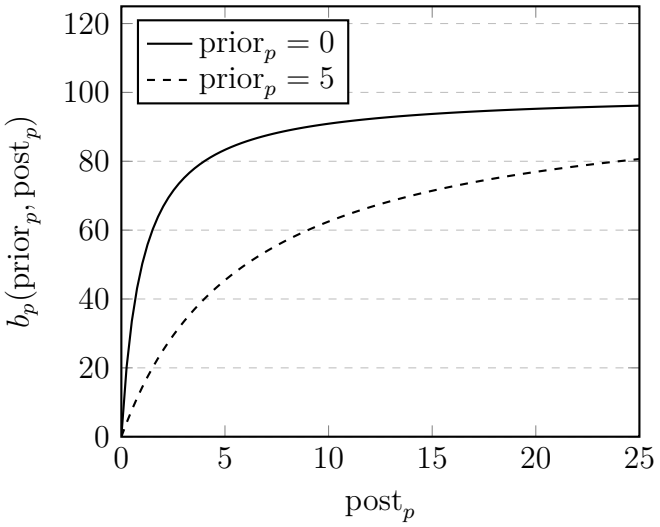
Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. "The Increasing Dominance of Teams in Production of Knowledge". In: *Science* 316.5827, pp. 1036–1039.

Xu, Fengli, Lingfei Wu, and James A Evans. "Flat teams drive scientific innovation". In: *PNAS* 119.23.

Additional Tables and Figures

Figure S1

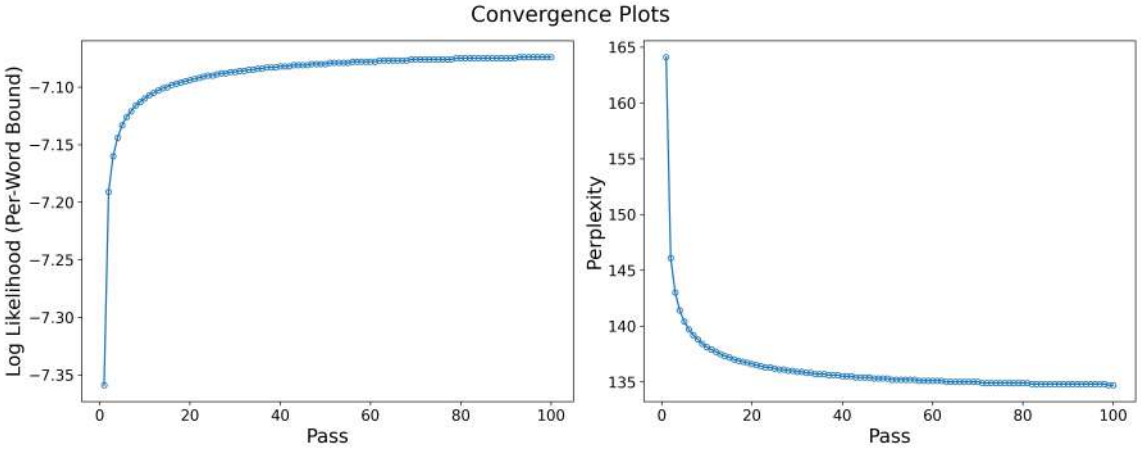
RAW BREAKTHROUGH MEASURE



Notes: Function for the raw breakthrough measure at the patent level. This measure is bounded between 0 and 1, but importantly captures a concept of percentage change even when the pre-count is equal to zero.

Figure S2

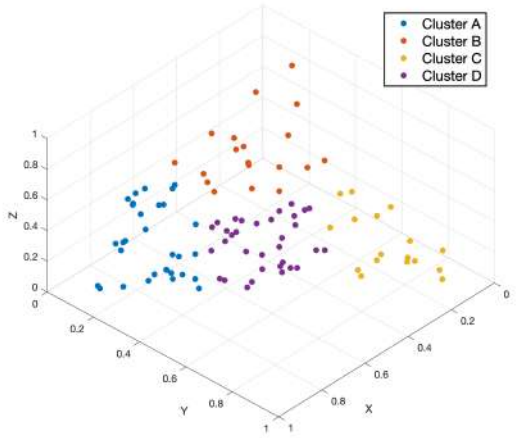
LDA CONVERGENCE



Notes: Convergence results, taken at the end of each of the 100 passes. For each pass the model slices the data into chunks of 2000 documents, and runs up to 350 iterations over these documents, or within-pass convergence.

Figure S3

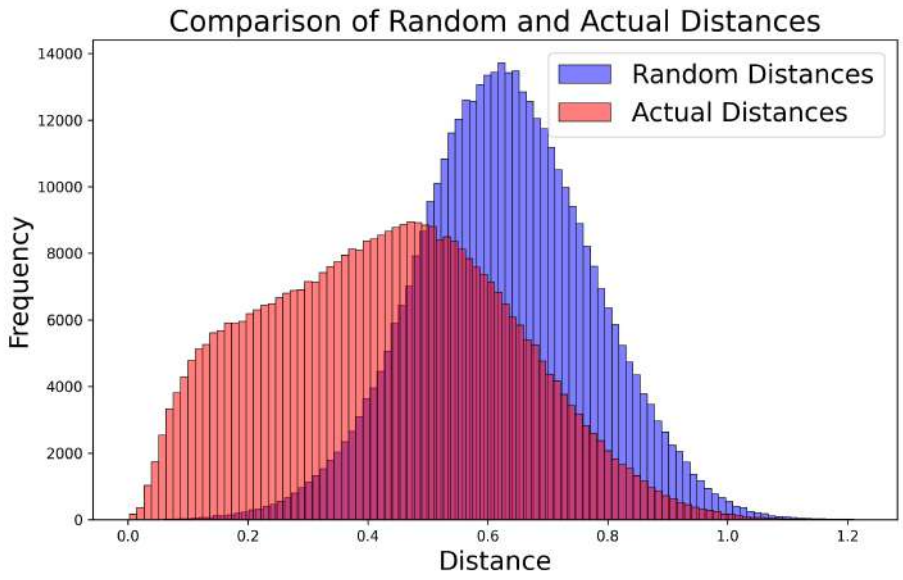
K MEANS CLUSTERING EXAMPLE



Notes: An example of how the k means clustering algorithm groups patents into broad knowledge groups. K-means clustering computed using the *sklearn KMeans* package.

Figure S4

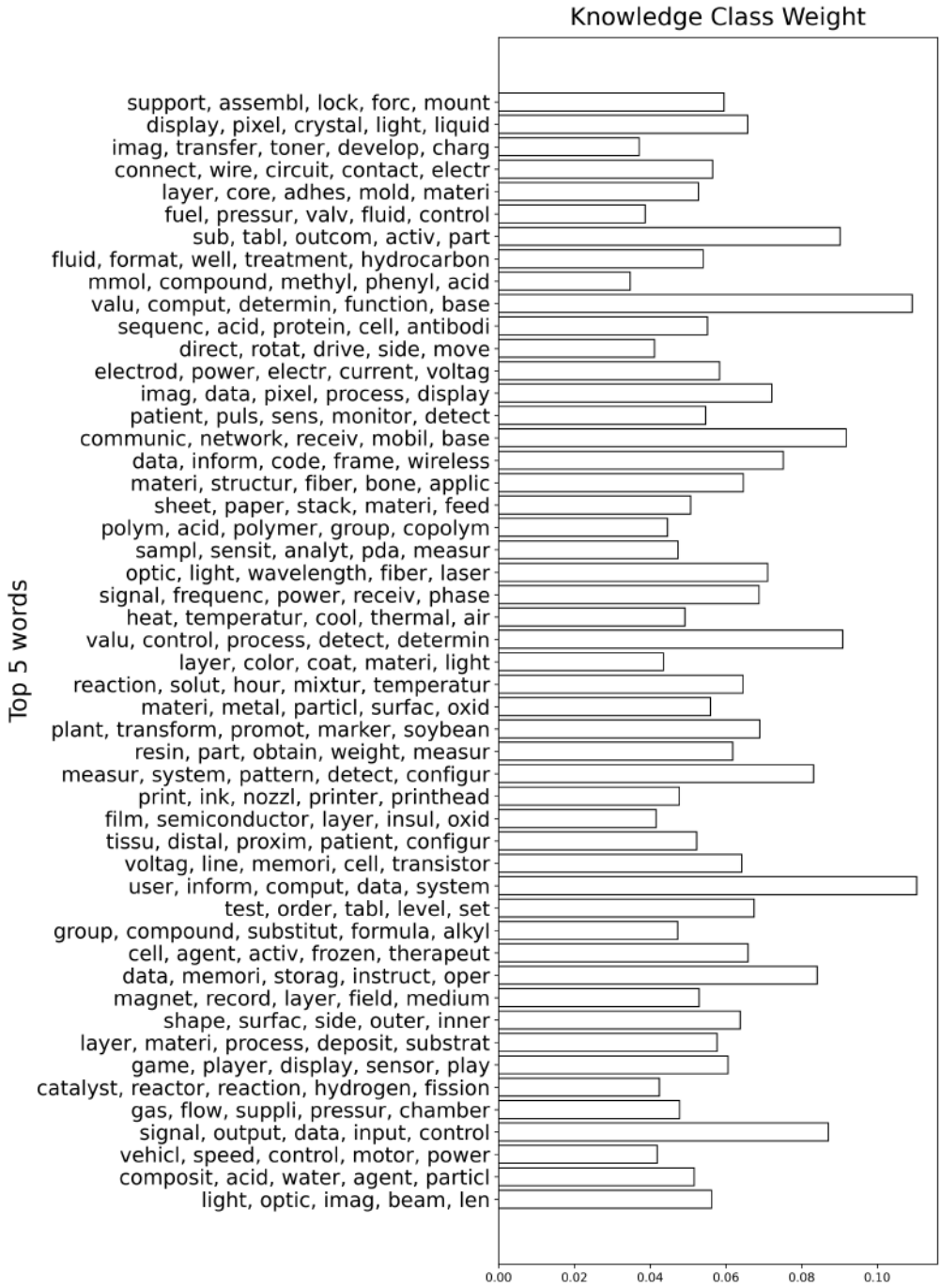
EXPECTED VERSUS REALISED PATENT DISTRIBUTION DISTANCES



Notes: This plots a histogram of the distances between the 2.2 million estimated patent distributions for the LDA sample and the realised patent distribution. and another draw uniformly at random.

Figure S5

INFERRED BAYESIAN PRIOR α



Notes: Learnt α Dirichlet prior. The Y axis presents the 5 words with the largest weight within the distribution for that class.

Figure S6

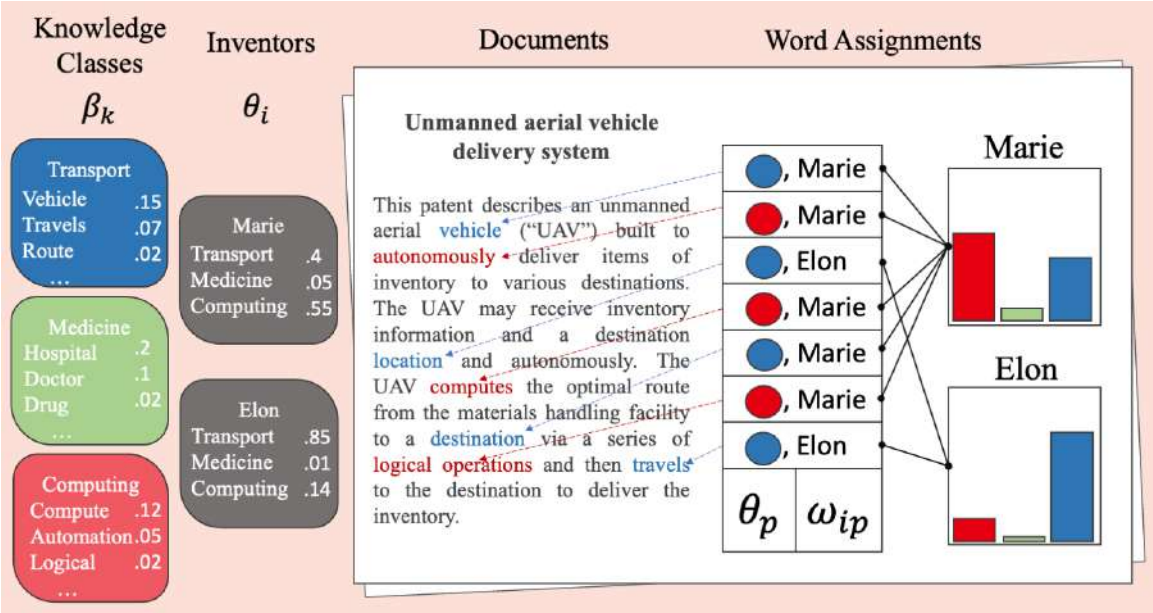
AGGREGATE TOPIC DISTRIBUTION BY PATENT TYPES

Notes: Patent knowledge distribution by type of patent. This is figure finds the average knowledge class weight, conditional on patent type.

Appendix

A Estimating the Knowledge Space: Intuitive example

Figure A1
INTUITIVE LDA EXAMPLE



Notes: An intuitive example of how LDA works. The example used is a paraphrased version of USPTO patent number US10839336B2 which expires 2036-06-12.

B Counting Objects in Knowledge Space

Recall that $n^i_{j(st)}$ denotes the count of j within i for s at t. To build count $n^A_{p'(pt)}$, the number of patents p' within the local knowledge field of a target patent p, it is straightforward to find all patents such that $\rho(\theta_p, \theta_{p'}) \leq r$. A patent p belongs to team span $S(\tau)$ if there exist a set of weakly positive weights that sum to 1 across the team member distributions to form a convex combination equal to the distribution for that patent.

To solve whether a patent p belongs to the local knowledge field of a team of n_τ members, I first find the closest point $\tilde{\theta} \in S(\tau)$ to that patent by finding the solution to the following problem.

$$\min_{\omega \in \mathbb{R}^n_\tau} \left\| \theta_p - \sum_{i \in \tau} \omega_i \theta_i \right\|$$
$$\sum_{i \in \tau} \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0$$

The objective is too choose the set of weights, such that they form a convex combination of each team members knowledge distribution, to minimise the distance between that point and the target patent distribution. If the distance between these two points is zero then this patent belongs to the convex hull of the team. If this distance is below the defined radius r, which remains constant across patents and teams, then this patent belongs to that teams local knowledge field.

I need to solve this problem for all patents in the sample, for each team. This is a huge number of problems to solve, in order to reduce the computational burden I take the following mathematical shortcut. I first calculate the centroid of the team span $S(\tau)$ as

$$c = \frac{1}{n_\tau} \sum_{i \in \tau} \theta_i$$

Calculate the maximum distance from the centroid to any point within the team vector using

$$d_{\max} = \max \|\theta - c\|$$

using the euclidean norm. For each patent θ_p calculate the distance between that patent distribution and the centroid $d = \|\theta_p - c\|$

Notice that any point which is further form the centroid than the maximum dis-

tance within the team span plus the radius r cannot form part of the local knowledge field. Therefore only solve the problem specified for those patents which

$$d_i \leq d_{\max} + r$$

Since this calculation is computationally far less demanding and faster than solving the problem, but ultimately gives the same solution.

C Technical Appendix: LDA

This technical appendix outlines the Latent Dirichlet Model (LDA) and the estiamtion process used. Modelling documents as a mixture of topics, where each topic is a distribution over words was brought into mainstream computer science by the LDA model presented in David M. Blei, Ng, and Jordan 2003. The Author-Topic-Model was first introduced by Rosen-Zvi et al. 2012. This is replication of the model in Mortensen 2017 where I have simply adapted the notation from the original papers to the context of an Inventor-Knowledge Class-Model, where inventors write patent texts collaboratively.

A patent p is a vector of N_p words $\boldsymbol{w}_p$ where each word w_{ip} is chosen from a vocabulary of size V, and a vector of n_τ inventors $\boldsymbol{i}_\tau$. A collection of P patents is therefore defined as $\mathcal{P} = \{(\boldsymbol{w}_1, \boldsymbol{i}_1, \cdots, (\boldsymbol{w}_P, \boldsymbol{i}_T)\}$.

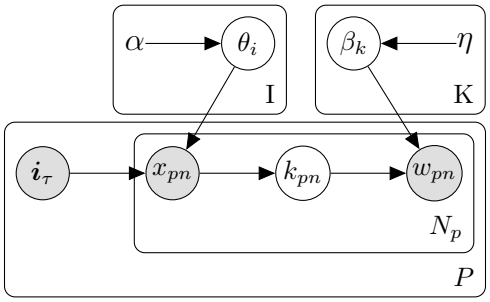
A set of patents is produced with the following generative process where the baseline assumption is that each inventor is drawn with uniform probability, such that from the law of large numbers, over sufficiently long patents each inventor contributes equally.

- For each inventor $i \in \{1, \ldots, I\}$ draw $\theta_i \sim \text{Dir}(\alpha)$.
- For each knowledge class $k \in \{1, \ldots, K\}$ draw $\beta_k \sim \text{Dir}(\eta)$.
- For each document $p \in \{1, \ldots, P\}$:
  - Given the team τ of patent p
  - For each word in the patent $n \in \{1, \ldots, N_p\}$.
    - Assign an inventor to the current word by drawing $x_{pn} \sim \text{Unif}\left(\frac{1}{n_\tau}\right)$.

- Conditioned on x_{pn}, assign a knowledge class by drawing $k_{pn} \sim \text{Mult}(\theta_i)$.
- Conditioned on z_{pn}, choose a word by drawing $w_{pn} \sim \text{Mult}(\beta_k)$.

This model is represented in the following plate diagram in figure (12).

Figure A2
INVENTOR-KNOWLEDGE CLASS MODEL



Notes: Plate notation for Bayesian Hierarchical model.

The posterior given the observed data and Dirichlet priors is given by

$$P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{w}, \alpha, \eta, \boldsymbol{T}) = \frac{P(\boldsymbol{w}|\boldsymbol{k}, \boldsymbol{\beta}) P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta}) P(\boldsymbol{i}|\boldsymbol{T}) P(\boldsymbol{\beta}|\eta) P(\boldsymbol{\Theta}|\alpha)}{P(\boldsymbol{w}|\alpha, \eta, \boldsymbol{T})} \tag{14}$$

As is typical in Bayesian analysis this posterior is intractable since we have no estimate for the marginal probability of the observed data. Therefore topic models typically use an inference method called Variational Bayes[12]. Define $q(\cdot)$ as an approximation to the posterior

$$q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \lambda, \gamma, \phi) = q(\boldsymbol{\Theta}|\boldsymbol{\gamma}) q(\boldsymbol{\beta}|\lambda) q(\boldsymbol{k}, \boldsymbol{i}|\phi) \tag{15}$$

$$\approx P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{w}, \alpha, \eta, \boldsymbol{T}) \tag{16}$$

Equation (3) models the knowledge classes and inventors as dependent random variables where $P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta}) P(\boldsymbol{i}|\boldsymbol{T}) \approx q(\boldsymbol{k}, \boldsymbol{i}|\phi)$. This is known in the literature as a blocking estimator. This means that the probability of choosing inventor $i \in \tau_p$ is a function of the knowledge held by inventor i relative to their collaborators, and the

[12]A derivative of Expectation Maximisation. Gibbs Sampling is an alternative and popular model, which can give good results and I have applied, however on large sample sizes can perform very slowly.

knowledge contained in the patent p. If a patent includes a lot of words discussing medicine, then if one of the inventors has a larger weight in this knowledge class than others in the team, they are more likely to be chosen to contribute. This allows for non-uniform contribution weights $\omega_{ip} \neq \omega_{jp} \ \forall \ i, j \in \tau$ and for the knowledge profile of individual inventors to be over(under) represented in the patent knowledge distribution.

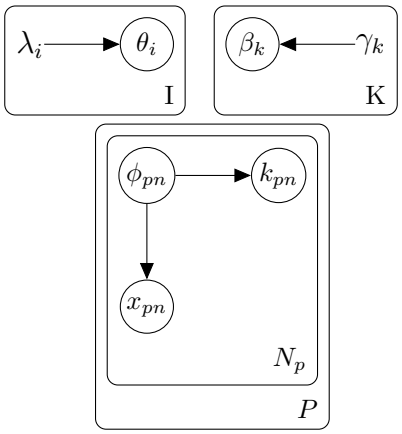
Define the following parametrisation of $q(\cdot)$

$$\begin{aligned} q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\lambda, \gamma, \phi) &= q(\boldsymbol{\Theta}|\boldsymbol{\gamma}) q(\boldsymbol{\beta}|\lambda) q(\boldsymbol{k}, \boldsymbol{i}|\phi) \\ &= \prod_i q(\theta_i|\gamma_i) \prod_k q(\beta_k|\lambda_k) \prod_{p,n} q(i_{pn}, k_{pn}|\phi_{ik}) \\ &= \prod_i Dir(\theta_i|\gamma_i) \prod_k Dir(\beta_k|\lambda_k) \prod_{p,n} q(i_{pn}, k_{pn}|\phi_{ik}) \end{aligned}$$

Which is the product of the probability of observing I individual knowledge class distributions, K knowledge class to word distributions and a set of inventor and knowledge class combinations for each word of every patent.

By changing the underlying assumption of how inventors and knowledge classes are drawn, to more closely match reality, the plate diagram of parameter dependence changes. Figure (13) presents the final model.

Figure A3
INVENTOR-KNOWLEDGE CLASS MODEL: BLOCKED



Notes: Plate notation for Bayesian Hierarchical model in a blocked model, given the assumption that the draw of inventor and knowledge class are dependent, thus allowing for non-uniform contribution shares.

For a given patent p the matrix ϕ_{ik} gives the discrete joint probability of choosing each inventor i and knowledge class k combination for a given word $n = v \in V$. Formally, the probability of inventor i choosing knowledge class k and word v for patent p is given by

$$\phi_{ivk} = \begin{cases} \phi_{ivk} & i \in \tau_p \\ 0, & \text{otherwise} \end{cases}$$

The full probability distribution is stored during the estimation as a four dimensional matrix ϕ_{pvik}[13]. Where $\sum_{i \in \tau} \sum_k \phi_{pvik} = 1$.

The model iterates over every word of each patent and updates the estimates for the parameters using the expected values. The method is a derivation of Expectation Maximisation and solves for the following condition using Jensen's inequality[14]

$$\begin{aligned} log\ p(w|\alpha, \eta, \boldsymbol{T}) \geq \\ log(\mathbb{E}_q[P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\alpha, \eta, \boldsymbol{T})]) - log(\mathbb{E}_q[q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\lambda, \gamma, \phi)]) \\ = \mathcal{L}(\lambda, \gamma, \phi) \end{aligned}$$

The right hand side is a lower bound on the marginal probability of the observed data. Also known in the literature as the Evidence Lower Bound (ELBO). Given the functional assumptions you can solve the right hand side by defining the expected values. The goal is then to maximise this right hand side as to approximate the log likelihood of the observed data as closely as possible. This is done through coordinate ascent, which maximises a multivariate function by iterating over each variable and optimising in that direction, holding all others constant until convergence. To do so take the derivative of $\mathcal{L}(\lambda, \gamma, \phi)$ with respect to the arguments to define three update rules, one for each variational parameter.

On convergence, I back out the θ_i given γ_i and β_k given λ_k. I do so using the process outlined in the literature so again, leave the interested reader to consult Mortensen 2017 for further details. The model presented here though, in addition to

[13]In reality the Gensim package uses the exchangeability of the model to develop an online algorithm to reduce the memory requirements of this matrix, I refer you again to Mortensen 2017 for further details on this great package.

[14]For a full derivation I refer the reader to the original paper by David M. Blei, Ng, and Jordan 2003

estimating a set of θ_i and β_k, estimates a contribution share for each team member and a set of patent to knowledge class distributions. To do so I sum across the relevant dimensions of ϕ_{pvik} as

$$\phi_{pvik} = \frac{\exp\{\mathbb{E}_q[\log\ \theta_{ik}] + \mathbb{E}_q[\log\ \beta_{kv}]\}}{\sum_k \sum_{i \in \tau_p} \exp\{\mathbb{E}_q[\log\ \theta_{ik}] + \mathbb{E}_q[\log\ \beta_{kv}]\}}$$

On convergence, the matrix ϕ_{pvik} is then given as part of the optimal solution. I then calculate the contribution shares and patent distributions in the following manner

$$\omega_{ip} = \sum_{vk} \phi_{pvik}$$
$$\theta_p = \sum_{i \in \tau} \omega_{ip} \theta_i$$

D Robustness Tests

Table A1
Treatment Team Regression Estimates: Adding

I: Breakthrough

	Dependent variable: Pr(Breakthrough)			
$D_{\tau'} \cdot \mathbb{1}$	-0.0133	-0.0161	-0.0020	-0.0039
	(-1.55)	(-1.70)	(-0.27)	(-0.42)
$n_{\tau t}$	-0.0320***	-0.0407***	-0.0116*	-0.0111
	(-5.27)	(-5.09)	(-1.97)	(-1.91)
Volume_{τ}				0.4377
				(0.42)
N	940	940	940	940
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

II: Direction

	Dependent variable: Pr(Direction)			
$D_{\tau' t}(\text{Direction}) \cdot \mathbb{1}$	0.0066*	0.0066*	0.0086**	0.0105***
	(2.50)	(2.48)	(2.98)	(3.33)
Prior work | Direction	0.0852***	0.0864***	0.0871***	0.0868***
	(22.72)	(22.83)	(22.79)	(22.72)
Volume_{τ}				-0.3870
				(-1.84)
N	7875	7875	7875	7875
Team FE	✓	✓	✓	✓
Period × Direction FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

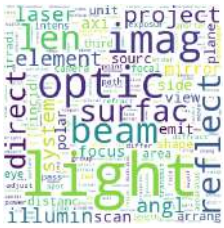
Notes: All regressions are team and patent order fixed effects models and standard errors

are clustered at this level. The dependent variable for panel I) is an indicator for whether the patent is a breakthrough using the Kelly et al. 2021 data. The dependent variable for panel II) is a stacked indicator for whether a patent achieves the given direction: mitigates climate change, reduces cancer risk or automates production. Controls include $d(\theta_p^e, \theta_p)$.

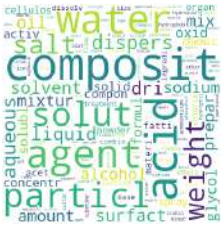
E Wordclouds

These are the fifty wordclouds, one for each of the estimated knowlegde classes, in addition to figure 9, here the relative size of each word in each knowledge class is visable.

1)



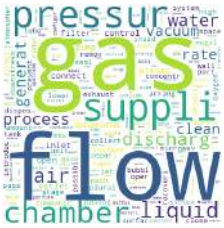
2)



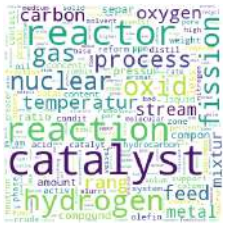
3)

4)

5)



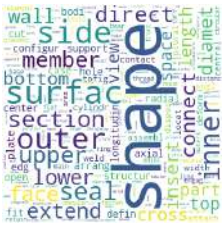
6)



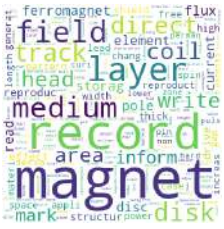
7)

8)

9)



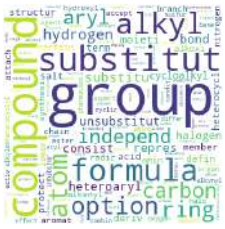
10)



11)

12)

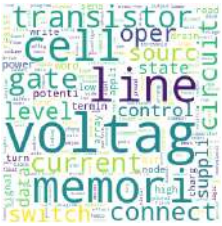
13)



14)

15)

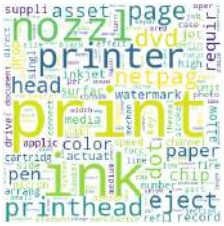
16)



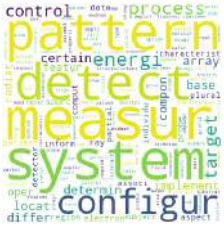
17)

18)

19)



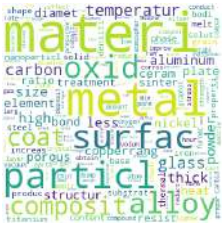
20)



21)

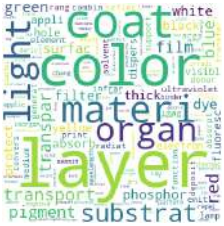
22)

23)



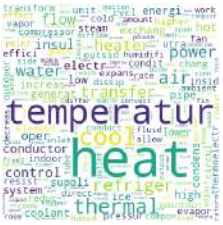
24)

25)

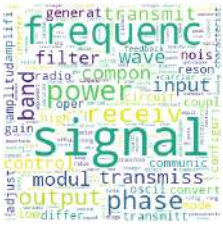


26)

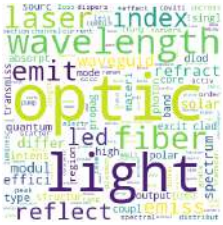
27)



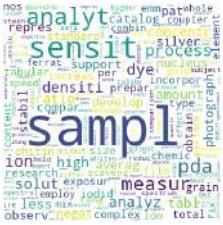
28)



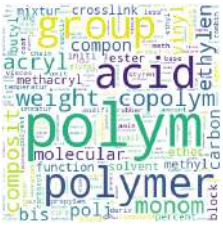
29)



30)



31)



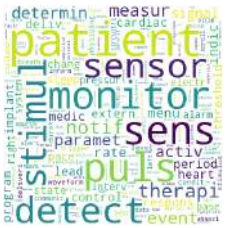
32)

33)

34)

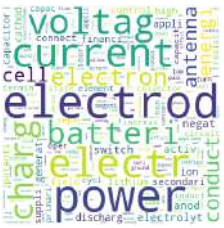
35)

36)



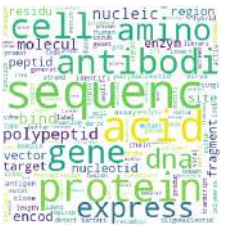
37)

38)

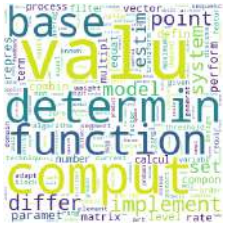


39)

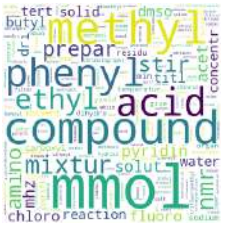
40)



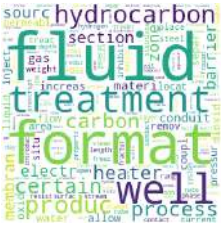
41)



42)



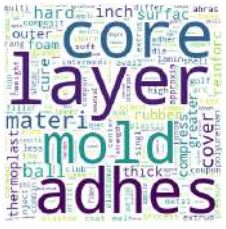
43)



44)

45)

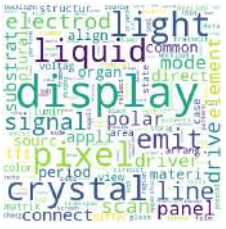
46)



47)

48)

49)



50)