Teams and Text: Collaborative Innovation in the Knowledge Space

Joseph Emmens[‡]

October 17, 2024

Click here for the latest version

Abstract

This paper introduces a novel framework that integrates inventor teams and their patent texts. The model locates inventors, teams, and research fields within an innovation ecosystem, referred to as the *knowledge space*. I use this framework to measure the impact of a research field's development stage on team outcomes. I study how the quantity and type of prior work available shape a team's ability to create new research fields and to direct existing ones toward specific targets. For targeted innovations, I examine automating production, mitigating climate change, and reducing cancer risks. The paper is the first to incorporate a Bayesian model of Natural Language Processing directly into a model of collaboration. The high-dimensional patent text data allows the derivation of an empirical measure for a team's local knowledge field. I combine this with data on premature inventor deaths as a quasi-natural experiment. I find that teams produce more breakthrough innovations when they reduce their size and move into under-explored research areas. However, when targeting a specific objective, teams benefit from adding a new member and incorporating existing knowledge from established fields.

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

Teams are the driving force behind advances in science and technology. Organising inventors into effective teams is essential for creating new research fields and addressing society's greatest challenges. According to endogenous growth theory, innovation arises from building on existing knowledge, and teams innovate by recombining the diverse expertise of their individual members (Weitzman, 1998). Yet, the knowledge production process remains largely a black box.

In this paper, I introduce a new framework that integrates inventor teams and their patent texts. The model locates inventors, teams, and research fields within an innovation ecosystem, referred to as the knowledge space. I use this framework to measure the impact of the development stage of a research field on a team's ability to innovate. Are teams more likely to create breakthrough innovations when building on a large body of existing work? Also, do teams direct their innovations toward a specific objective when there are many previous patents targeting the same outcome?

To answer these questions, I use a Bayesian model of Natural Language Processing (NLP) to open the black-box around teamwork. This model allows me to move beyond observing patent outcomes and understand the mechanism of how teams innovate. The paper is the first to incorporate a Bayesian NLP framework directly into a model of collaboration. The high dimensionality of unstructured patent text data allows the model to disentangle individual contributions and infer a team's local knowledge field. This is the set of all research fields within the knowledge space in which a team could innovate. I train the model on 408,841 patents from 214,535 teams using the USPTO *PatentsView* database. I then use the premature death of a team member (Kaltenberg et al., 2021) as a quasi-natural experiment. This approach identifies how changes to a team's knowledge field affect their ability to innovate along two dimensions. Their ability to produce breakthrough innovations (Kelly et al., 2021) or to target innovations at automation (Mann and Püttmann, 2023), climate change mitigation (PatentsView, 2024), and cancer risk reduction (USPTO, 2024).

I document that over time, research fields have become increasingly crowded. This causes teams to produce fewer breakthrough patents. Notably, a team's first patent is the most likely to be a breakthrough. This suggests a hit-or-miss innovation environment where impact hinges on the success of the initial idea. For teams to produce new scientific breakthroughs, they should change research fields. I find that

following the premature death of a team member, the remaining inventors are more likely to produce a breakthrough if they move into under-explored research fields.

Unlike with breakthrough innovations, I find that prior examples improve a team's ability to target specific innovation goals. For instance, consider a team whose objective is to produce a self-driving car. If someone has previously combined automation technologies with other vehicles, they can build on this existing knowledge. Moreover, teams continue to target a given direction if their previous patents did so, demonstrating path dependency in innovation. As a result, when a team loses a member who contributed crucial knowledge from fields relevant to a specific direction, the team's likelihood of patenting in that direction significantly decreases. I confirm the robustness of this finding by showing that the effect reverses when a team gains a member instead of losing one.

I develop and apply a two-part empirical strategy to demonstrate the results. First, I quantify the amount and type of prior work a team builds on by backing out a latent representation of their local knowledge field from patent texts. I model the knowledge space as a probability simplex across a set of knowledge classes. Each class represents a specific domain of expertise. For example, computer science, biology, or graphic design. Both inventors and patents are characterised by their position in this space. Teams innovate by recombining the knowledge profile of each member. A team's local knowledge field is then defined as all possible combinations of its members. Through a simplex, this approach naturally incorporates a spatial concept by embedding a notion of distance. This lets me measure the development stage of research fields available to the team by counting the amount of prior work in their knowledge area.

A premature death, defined as the death of an active collaborator, serves as a random shock to a team's local knowledge field. The use of premature deaths as a source of exogeneity is well established (Azoulay, Fons-Rosen, et al., 2019; Azoulay, Graff Zivin, et al., 2010). The death of a collaborator changes the set of research fields available to the team by removing potential combinations. I apply a continuous treatment model to show that a team's innovation output depends on the amount and type of prior work in this new set of research fields. The impact of a death on a team's research potential depends on which team member is lost and their contribution to the team's local knowledge field. The average treatment is the mean decrease in the quantity of prior work in a team's local field after a premature death. I can

predict how this change influences a team's ability to innovate. Following such a treatment, the likelihood of producing a breakthrough increases by 6.6 percentage points, a 73.33% increase on the baseline. However, when conditioning on patents being a specific type, the probability of targeting that direction falls by 4.4 percentage points, a 23.53% reduction on the baseline.

The results can be understood through the lens of an endogenous growth paradigm. Prior work has opposing effects on team output: it acts as a barrier for breakthrough innovations but serves as a catalyst for targeted innovations. The impact depends on whether a team aims to create entirely new research fields or to redirect existing ones. For targeted innovations, prior work lowers the cost by providing a foundation, making it easier for teams to combine knowledge fields and acheive a given purpose. This aligns with the idea that moving up the quality ladder of development reduces implementation costs (Grossman and Helpman, 1991).

However, when the goal is to create a new research field, prior work becomes an obstacle. It not only prevents teams from being the first to develop an object but also establishes paradigms that shape future work. This relates to the literature on the burden of knowledge (Jones, 2009), which suggests that as the frontier of knowledge expands, inventors must invest more effort to develop on that frontier. At an aggregate level, as the knowledge space fills up, breakthrough ideas become increasingly difficult to create (Bloom et al., 2020).

These findings provide guidance for policymakers. While established fields offer a strong foundation for targeted innovation, they can limit breakthroughs. Diversifying funding into newer fields could help teams combine ideas in novel ways and foster breakthroughs. For targeted goals like automation or mitigating climate change, researchers should bringing in expertise from well-developed fields and promote the use of cross-field collaboration.

On a technical level, this paper makes an important contribution to the use of NLP models in economics. This is the first paper to model innovation and the patent writing as one unified process. This provides an economic interpretation for the parameters inferred from the text analysis. 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.

As a novel method, I validate this space along various dimensions by comparing the model to existing data. I validate the contribution share for each inventor 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 there is a large gap in contributions, the lead inventor's past patents provide, on average, 14% more predictive information. This difference disappears as their contributions become more equal. In addition, the model develops a measure of 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.

Related Literature The first broad 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 et al., 2007). A range of reduced form papers have looked to describe team composition and its role in explaining innovation outcomes. Team size, hierarchy and diversity are key determinants (Uzzi et al., 2013; Xu et al., 2013; Wu et al., 2019). 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.

There has been significant interest in explaining complementarities between team members' knowledge and skills (Boerma et al., 2021; Freund, 2022; Pearce, 2022; Herkenhoff et al., 2024). This literature has made important contributions to understand how teams create value. However to study how teams create new research fields

[1]Using the data kindly provided online by Arts et al., 2021

and re-direct existing ones, a new framework was required. This paper proposes an alternative knowledge production process to the dominant CES production function. As the first paper to take a model of collaboration to data on patent texts, I am able to extend this literature to study breakthroughs and direction.

I contribute specifically to a smaller empirical literature which looks to disentangle individual contributions to team projects. A key feature of this paper is to allow team members to contribute in different amounts to the knowledge contained in an 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 Jones, 2019). There is a small but important literature using highly specific case studies in which individual inputs are observed (Kahane et al., 2013; Devereux, 2018; Weidmann and Deming, 2021). Alternatively, Bonhomme (2022) combines a model of team formation with a fixed effect method, to identify the marginal contribution of a team member to the quality of team output.

Finally, the third literature I contribute to is the use of natural language processing models within social science studies. This paper employs a two-stage approach to back out the required latent variables from text, as in Bandiera et al. (2020).[2] The closest paper to mine is the forthcoming study from Teodoridis et al. (2022). They develop a version of an LDA model, a Hierarchical Dirichlet Process (HDP) at the patent level to map the knowledge space over time. I extend this literature to the team level to study the production of knowledge.

Arts et al. (2021) developed the literature beyond using citations histories 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 defines the two stage empirical strategy. The full

[2]This paper therefore is susceptible to the subsequent critique raised in Battaglia et al. (2024). As cited in the original paper, this critique is relatively minor when using patent texts due to their length, and therefore wealth of data.

treatment of the LDA method used can be found in the technical appendix. Section 4 presents a set of descriptive statistics and validation tests. Section 5 presents the main results. Section 6 concludes.

2 Theoretical Framework

Define $\mathcal{K}$ as a set of K knowledge classes.[3] 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 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

[3] 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 Blei and Lafferty, 2005 for further details.

element in the support of a Dirichlet distribution can be treated as a K-dimensional discrete probability distribution.[4]

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.[5] 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.[6] 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. The team first draws the number of words in the patent $N_p \sim G(\cdot)$.[7] 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.

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

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

[6]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.

[7]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.

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

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

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

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 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.

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 belong-

ing to the local knowledge field can vary over time.

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. This example is informative as can be plotted in 2-D, and while the number of classes is small, there number of combinations remains infinite.

I define $S(\tau)$ as the team span: the set of all linear combinations of the team members' knowledge distributions. 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 $\tilde{S}(\tau)$.

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.

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 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.[8]

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

$\mathbb{1}$ denotes an indicator function which is equal to one when the condition in paren-

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

theses 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 decrease in the density of patents within a team's local knowledge field increases the probability that their next patent sparks a breakthrough.*

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 (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 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 $\partial f_z / \partial n_{pt}(z) > 0$, an increase in the density of patents of type z within a team's local knowledge field increases the probability that their next patent targets direction z.*

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 infer 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 build the sample around three types of teams, which I later combine into 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 et al. (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 m 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 extract the full patenting history of each member of every team. I train the LDA on this sample. 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.[9] 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).

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.*

[9] 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

This quote demonstrates that the patent texts are informative on the knowledge content beyond a simple title or CPC classification. The text describes features of the innovation that can be applied to other fields. 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 Blei, Ng, et al., 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, most importantly: 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 machine learning 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 patent technology classes evenly across all team members. Therefore in this case the method presented here defaults to the standard method in the literature (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. For more details consult the *atm* package documentation online.

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 et al. (2018) and 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.[10] Figure S2 plots the log-likelihood and perplexity at each pass over the data which shows that the model converges after approximately 100 passes. 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.

[10]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.

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 8 where the clusters form four broad groups of patents.

3.4 Empirical Strategy

I present a set of regression models to test both hypotheses derived in section 2. This paper therefore employs a two-stage method, first inferring a set of latent parameters, and then studying how they explain data in a reduced form model. 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 probability 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 11. The regression is run as a logit given

that both outcomes are latent variables. I model the probability of each outcome $y_{t\tau(p)} \in \{b_p, z_p\}$ as

$$Pr\Big(y_{t\tau(p)} = 1 \mid X'_{t\tau(p)}\boldsymbol{\psi}\Big) = \frac{\exp\Big(X'_{t\tau(p)}\boldsymbol{\psi}\Big)}{1 + \exp\Big(X'_{t\tau(p)}\boldsymbol{\psi}\Big)} \tag{10}$$

where

$$X'_{t\tau(p)}\boldsymbol{\psi} = \beta_0 + \beta_1 x_{t(p)} + \beta_2 d_p + \beta_3 \cdot t + \beta_4 m_\tau + \delta_c + \delta_z \tag{11}$$

For $x_{t(p)} \in \{n_{t(p)}, n_{t(p)}(z)\}$. Either the count of all patents belonging to the local knowledge field of patent p prior to their publication in time t, 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 $d_p = d(\theta_p^e, \theta_p)$ as defined in equation 2. I include the team size as a control with m_τ. I then include a set of fixed effects for knowledge cluster (c) as defined in section 3.3 and, when estimating the direction model, direction (z). Introducing a time trend through β_3 achieves multiple objectives. 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 et al. (2021). All variables are defined analogously as in equation 11, 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 and count those that belong to the local knowledge field of team τ.

Equation (13) measures changes in team output, holding team composition fixed. Here the outcome variable varies for teams across patents and I explain the probability that patent p from team τ achieves outcome y such that the logit model is now given

by

$$Pr\Big(y_{\tau p} = 1 \mid X'_{\tau p}\boldsymbol{\psi}\Big) = \frac{\exp\Big(X'_{\tau p}\boldsymbol{\psi}\Big)}{1 + \exp\Big(X'_{\tau p}\boldsymbol{\psi}\Big)} \tag{12}$$

where

$$X'_{\tau p}\boldsymbol{\psi} = \alpha_\tau + \delta_{sz} + \beta_1 n_{\tau t} + \beta_2 d_p + \beta_3 \cdot t + \delta_c \tag{13}$$

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_{sz} corresponds to the patent order (1,2,3 etc.) in each direction. Therefore the panel is defined at the team, patent order level (τ, s). 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. I include the time trend to control for systematic changes over time, as previously discussed. 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 first team, denoted τ_1 returns to patent within 5 years denoted as τ_2. The identifier τ is now a unique id for each pair (τ_1, τ_2). Either τ_2 consists of the original team minus the deceased inventor ($\tau_2 = \tau_1 \setminus \{i\}$), or they replace i with one other inventor j ($\tau_2 = (\tau_1 \setminus \{i\}) \cup \{j\}$).

I define the measure $D_{\tau t} = n_{\tau_1 t} - n_{\tau_2 t}$ to measure the change in the quantity of prior work on which the team is building, following their shift in the knowledge space.[11] $D_{\tau t}(z)$ is analogous but only counts those patents targeting direction z. I control in the regression for the first teams count $n_{\tau_1 t}$, 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 local knowledge field. Therefore I introduce a control for the volume denoted v_τ. I

[11] To ensure the logit model converges, I winsorize the top 1% of both the total and direction counts, $n_{\tau t}$ and $n_{\tau t}(z)$ respectively. This caps the maximum value at the 99% value, to reduce the effect of outliers.

approximate the volume of a team local knowledge field using the following[12]

$$v_\tau = \sqrt{m} \times (\alpha \cdot D_{\text{max}} + (1-\alpha) \cdot D_{\text{mean}})) = \text{vol}\Big(S(\tau)\Big) \tag{14}$$

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. This is not a function of r since that is constant across teams. Again this is run on the team, patent panel such that the logit model is as specified in equation 12 but I extend the independent variables.

$$X'_{\tau p}\boldsymbol{\psi} = \alpha_\tau + \delta_{sz} + \beta_1 D_{\tau t} \cdot \mathbb{1}_{post} + \beta_2 v_\tau + \beta_3 n_{\tau_1 t} + \beta_4 d_p + \beta_5 \cdot t + \delta_c \tag{15}$$

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 the results to data taken from the literature.

Table 2 describes the sample used in the continuous treatment model. This sample is defined as in section 3.1, around a set of treated teams. Therefore this is not the full sample on which the LDA was trained.

This paper examines how the maturity of a research field determines the innovation outcomes for teams working in that area. While the knowledge profile and team span is constant over time, innovation arrives dynamically to the knowledge space. Therefore the research field in which a team works develops over time.

[12]Measuring the volume in high dimensional space is challenging, and there are alternative ways to do this. For example, using a set of uniformly distributed points and sampling via MCMC to order team span sizes.

Table 2
Descriptive Statistics

	Treatment Status			
	No Replace	Replace	Control	
N = 9009	353	2328	6418	
Full sample	1990-2010			
	Obs	Mean	Min	Max
Team size	9009	2.505	1	20
Volume	9009	0.725	0	4.69
Total Count	9009	269.41	0	3048
Density	7018	426.27	0	58353
Patents	9009	3.995	0	53
% Green	9009	0.055	0	1
% Automation	9009	0.46	0	1
% Cancer	9009	0.03	0	1
% Breakthrough	8304	0.09	0	1

Notes: Descriptive statistics for the N = 9009 teams in the estimation sample of the continuous treatment model. This model uses a team × patent panel. Volume is defined in equation 14 and total count redefines equation 8 to sum over the team local field. Density is defined as total count divded by the volume, and is set to missing as some teams of size 1 have zero volume and zero count.

According to Bloom et al. (2020), innovative ideas are getting harder to find. This in part can be explained by an increasingly populated knowledge space. Figure 3 plots the average density of a team's local knowledge field across the sample, for teams which patent for the first time in each year. The density of prior work within a team's local knowledge field is defined as $n_{\tau t}/v_{\tau}$.

As the innovation frontier expands, the number of patents within the knowledge space is constantly increasing. This doesn't however mean that the number of patents within team local knowledge fields increases, as teams may endogenously respond and locate themselves in less populated areas. However, I find that when teams produce their first patent, the number of prior work within their field on average is increasing, as is the density.

This is driven in part by the fact that the number of patents within team's local fields is increasing over time, while the volume of team fields is relatively constant. This is an interesting result, given that team size is increasing. This however is compensated by teams combining inventors who are closer together, such that the

maximum and mean distance between members is decreasing. The comparison across team statistics and the breakdown of the volume measure can be seen in Figure S7.

Figure 3
Average Knowledge Field Density

Notes: Density is defined at the team level, as the number of patents within their local knowledge field normalised by their volume. I then find the average density for each year, of team's which patented for the first time in that year. The size of each marker is weighted by the total number of patents which were applied for (and later granted) in each year.

Figure 3 shows that on average team's local knowledge fields are filling up. This however could be an endogenous choice of the teams to locate themselves on increasingly developed areas. This justifies the identification strategy around premature inventor deaths. The objective is to find exogenous variation in the amount and type of prior work a team builds on. This will allow me to make conclusions on the impact of increasingly dense local knowledge fields on team output.

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. I will compare the patents identified as breakthroughs by equation 16 to the literature to demonstrate the precision of this method.

$$b_p = \left(\frac{\text{post-count}_p}{1 + \text{prior-count}_p + \text{post-count}_p} \right) \times 100 \tag{16}$$

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 16 is the raw breakthrough measure, however as made clear in Hall et al. (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.

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 (4) 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 of residuals from a regression of b_p on a set of application year dummies.

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 3 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 et al. (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 classifications	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)	%Δ new re-used words in breakthrough patents	8.67^{***}
	%Δ new re-used bi-grams in breakthrough patents	47.6^{***}
	%Δ new re-used tri-grams in breakthrough patents	44.1^{***}
Citations	%Δ forward citations for + Δ1% in post count$_p$	$2.07\%^{***}$
	%Δ backward citations for + Δ1% in prior count$_p$	$1.09\%^{***}$
	Δ% forward citations for breakthrough patents	$16.2\%^{***}$

Notes: Validation statistics using UPSTO citation data and existing patent novelty literature. The correlation between the Arts et al. (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.

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.

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.

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 4

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 4 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 in section 2 using the empirical strategy detailed in section 3.4. I treat the breakthrough and direction models separately. In the breakthrough model, the outcome variable is taken from

Kelly et al. (2021).[13] 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 16.

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. In Table 5 panel A) 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. This is to be expected, when patenting in well-researched areas, there is less room to produce a breakthrough.

In Table 6 panel A) I 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 around 0.01%. This is partly driven by the teams own innovation. A team's first patent is most likely to be a breakthrough, since they develop on their initial idea, each new patent is decreasingly likely to spark a new field.

Finally, Table 7 reports the coefficient of interest on the response of a team's patents to changes in the team's local knowledge field. 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 et al. (2019)). In this case small refers to specialised, as teams that reduce the distance between their members are more likely to produce breakthroughs.

For the direction model, 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, again defined in section 3.1. These three are not mutually exclusive. Recall that this corresponds to a stacked regression, controlling for each direction through a direction times period fixed effect.

[13] 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 5
PATENT REGRESSION ESTIMATES

PANEL A: 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				✓

PANEL B: 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 A) the dependent variable is the probability that patent is in the top 90% in the Kelly et al. (2021), for 5 years. In panel B) 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 $d(\theta_p^e, \theta_p)$.

From Table 5 panel B) 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. This demonstrates that the cost of innovating is reduced by prior work. The coefficient is an order of magnitude larger than the breakthrough model.

Table 6
WITHIN TEAM REGRESSION ESTIMATES

PANEL A: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
Prior work$_{\tau t}$	-0.0248***	-0.0239***	-0.0246***	-0.0103***
	(-15.34)	(-15.02)	(-15.25)	(-8.07)
N	6275	6275	6275	6275
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls		✓	✓	✓
Cluster FE			✓	✓
Time trend				✓

PANEL B: DIRECTION

	Dependent variable: Pr(Direction)			
Prior work | Direction$_{\tau t}$	0.0578***	0.0578***	0.0579***	0.0586***
	(67.81)	(67.82)	(67.86)	(67.99)
N	69480	69474	69474	69474
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 A) is a binary indicator for whether the that patent is a breakthrough, or not. The dependent variable for panel B) 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 6 panel B), I present supporting evidence for Hypothesis 2 without changing team composition. Each additional patent which arrives to the team's local knowledge field is associated with an increase in the odds that their next patent targets direction z of approximately 0.058%. This again can be in part driven by the team's own patenting. If the team first patents towards a given direction, the theory of path dependency states that their next patents are increasingly likely to do so. If their patent reduces the cost of others also patenting in that direction, neighbouring teams are also increasingly likely to patent in that direction x (Aghion et al., 2016).

Table 7

Treatment Team Regression Estimates

Panel A: Breakthrough

	Dependent variable: Pr(Breakthrough)			
$D_{\tau t} \cdot \mathbb{1}$	0.0057***	0.0053***	0.0030***	0.0024**
	(6.61)	(5.53)	(3.50)	(2.82)
Prior work$_{\tau_1 t}$	-0.0209***	-0.0218***	-0.0090***	-0.0089***
	(-15.52)	(-15.71)	(-8.35)	(-8.25)
Volume$_{\tau}$				-0.9976*
				(-2.57)
N	6920	6920	6920	6920
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

Panel B: Direction

	Dependent variable: Pr(Direction)			
$D_{\tau t}$(Direction) $\cdot \mathbb{1}$	-0.0138***	-0.0141***	-0.0139***	-0.0146***
	(-11.34)	(-11.54)	(-11.47)	(-11.66)
Prior work | Direction$_{\tau_1 t}$	0.0524***	0.0525***	0.0533***	0.0533***
	(68.83)	(68.90)	(69.05)	(69.07)
Volume$_{\tau}$				-0.3788**
				(-2.68)
N	71784	71784	71784	71784
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 identifier τ is unique for each pair (τ_1, τ_2). The dependent variable for panel A) is an indicator for whether the patent is a breakthrough using the Kelly et al. (2021) data. The dependent variable for panel B) 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 the treatment model confirms Hypothesis 2 in the data. For each additional patent targeting direction z that is lost from the team knowledge field, the odds of their next patent targeting that direction decrease by 0.0146%. When the team loses the relevant expertise, they can no longer innovate on those research fields that have a long history in a given direction.

To put these coefficients into tangible numbers consider the following comparison. Each inventor contributes differently to the team. The justification for a continuous treatment model is that it matters who is lost, and which knowledge they contributed to the team. The average treatment measures the typical impact on a team's local field when a team member is lost. By estimating the average number of patents typically lost after such an event, I can predict how this change influences a team's ability to innovate.

For the sample of teams which return to patent without replacing the deceased inventor, the average number of patents lost from a team's local knowledge is 261. The change in probability is calculated using a baseline probability of 0.09 taken from Table 2. The coefficient of 0.0024 from Table 7, panel A) column 4, and an average treatment of 261 yield a change in log-odds of 0.6264. Applying this to the baseline log-odds of -2.313 results in new log-odds of -1.6866. Converting this back to probability gives 0.156, indicating a change of approximately 6.6 percentage points.

The same calculations for the direction model give the following. The change in probability is calculated using a baseline probability of 0.187. This is the average across each direction. The coefficient of -0.0146 again from Table 7, panel B), column 4 and an average treatment of 22 yield a change in log-odds of -0.3212. Applying this to the baseline log-odds of -1.47 results in new log-odds of -1.7912. Converting this back to probability gives 0.143, indicating a change of approximately -4.4 percentage points.

This is for the case in which the inventor is not replaced, however in the majority of cases they are replaced. This choice is therefore endogenous to the decision of the team on the type of innovation they want to produce. In fact, the average treatment is positive for the teams which replace the inventor. This means that they replace the deceased inventor with another from a more established research field. The continuous model is therefore required to use variation in who the team chooses to replace the deceased inventor with.

5.1 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 Appendix section D. As a robustness check, instead of removing a team member I select a sample of teams that add one. In this case the coefficient sign should flip. Table 9 confirms that the results are robust to adding an inventor.

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 (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.[14] I introduce into equation 15 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 et al. (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 8 indicates that team members potentially learn from each other. Interestingly, for the breakthrough model, the interaction term is insignificant. While for the direction model, there is stronger support for the idea that inventors learn from each other. This appears logical, since knowledge specific to a direction can be learnt. While the ability to produce radical ideas is less tangible, and therefore harder to learn. This result should be taken cautiously. As one of many examples of

[14] I would like to thank Professor Florenta Teodoridis for their valuable suggestion to include this extension, which has greatly improved the clarity and depth of this paper.

the potential for future work, a dynamic version of the model would allow for a much more rigorous treatment.

Table 8
Learning Effects

A) Breakthrough		B) Direction	
	Pr(Breakthrough)		Pr(Direction)
$D_{\tau t} \cdot \mathbb{1}$	0.0050*	$D_{\tau t}(\text{Direction}) \cdot \mathbb{1}$	-0.0208***
	(2.52)		(-11.75)
$D_{\tau t} \cdot \mathbb{1} \cdot p_\tau$	-0.0012	$D_{\tau t}(\text{Direction}) \cdot \mathbb{1} \cdot p_\tau$	0.0033***
	(-1.46)		(5.25)
Prior work$_{\tau_1 t}$	-0.0089***	Prior work | Direction$_{\tau_1 t}$	0.0535***
	(-8.21)		(69.08)
Volume$_\tau$	-0.9990*	Volume$_\tau$	-0.3938**
	(-2.54)		(-2.78)
N	6920	N	71784
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 A) is a binary indicator for whether the that patent is a breakthrough, or not. The dependent variable for panel B) 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.

6 Concluding Remarks

In this paper I ask how the development stage of a research field determines innovation outcomes for teams working on that area. In order to answer this question I propose a new innovation production process. To do so, I combine a model of collaboration

and a Bayesian model of Natural Language Processing. This paper is the first to integrate a model of text analysis directly into a model of research collaboration. The paper builds a mapping of inventors, teams and patents in which to study how teams innovate. 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 allows me to tackle a set of important hypotheses on which systematic evidence was missing.

The framework developed in this paper is required to back out a latent representation of a team's local knowledge field. The combination of the high-dimensionality of patent text data, and the computational Bayesian model allows me to empirically model the set of research fields a team could work on. I then shock the local knowledge field of a team with the premature death of a collaborator.

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 changing team composition, the team can increase their probability of producing a breakthrough or innovations targeted at certain direction.

The findings highlight key strategies for policymakers. If they want to target a specific strategic outcome, like automation and climate change solutions, innovation is best supported by drawing expertise from mature fields through cross-field collaboration. However, while established fields provide a strong base, they may also limit opportunities for breakthroughs. To counter this, diversifying investment into emerging fields could enable teams to experiment with novel idea combinations, increasing the chances of breakthrough innovations.

The framework presented here opens a world of future research. The knowledge space provides a rich environment in which to study teams, but can be integrated with economic models to explain the broader innovation landscape. For example, modelling public R&D financing or firm innovation choices. Another key avenue for future work is to study the role of learning in this context and develop a dynamic version of the model. 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

Aghion, **Philippe** et al. (2016). "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry". *Journal of Political Economy* 124.1, pp. 1–51.

Ahmadpoor, **Mohammad** and **Benjamin F. Jones** (2019). "Decoding team and individual impact in science and invention". *PNAS* 116.28, pp. 13885–13890.

Akcigit, **Ufuk** et al. (2018). *Dancing with the Stars: Innovation Through Interactions*. Working Paper 24466. National Bureau of Economic Research.

Arts, **Sam**, **Jianan Hou**, and **Juan Carlos Gomez** (2021). "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures". *Research Policy* 50.2, p. 104144.

Azoulay, **Pierre**, **Christian Fons-Rosen**, and **Joshua S. Graff Zivin** (2019). "Does Science Advance One Funeral at a Time?" *American Economic Review* 109.8, pp. 2889–2920.

Azoulay, **Pierre**, **Joshua S. Graff Zivin**, and **Jialan Wang** (May 2010). "Superstar Extinction". *The Quarterly Journal of Economics* 125.2, pp. 549–589.

Bandiera, **Oriana** et al. (2020). "CEO Behavior and Firm Performance". *Journal of Political Economy* 128.4, pp. 1325–1369.

Battaglia, **Laura** et al. (2024). "Inference for Regression with Variables Generated from Unstructured Data". *Unpublished manuscript*.

Blei, **David M** and **John D Lafferty** (2005). "Correlated topic models". *NeurIPS*, pp. 147–154.

— (2006). "Dynamic topic models". *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120.

Blei, **David M.**, **Andrew Y. Ng**, and **Michael I. Jordan** (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3, pp. 993–1022.

Bloom, **Nicholas** et al. (2020). "Are Ideas Getting Harder to Find?" *American Economic Review* 110.4, pp. 1104–44.

Boerma, **Job**, **Aleh Tsyvinski**, and **Alexander P Zimin** (2021). *Sorting with Team Formation*. Working Paper 29290. National Bureau of Economic Research.

Bonhomme, **Stéphane** (2022). "Teams: Heterogeneity, Sorting, and Complementarity". *Unpublished manuscript*.

Devereux, **Kevin** (2018). *Identifying the value of teamwork: Application to professional tennis*. Working Paper Series 14. University of Waterloo.

Fleming, **Lee** and **Olav Sorenson** (2004). "Science as a map in technological search". *Strategic Management Journal* 25.8-9, pp. 909–928.

Freund, **Lukas** (2022). *Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities*. Working Paper. SSRN.

Griffiths, **Thomas L.** and **Mark Steyvers** (2004). "Finding Scientific Topics". *Proceedings of the National Academy of Sciences* 101.1, pp. 5228–5235.

Grossman, **Gene M** and **Elhanan Helpman** (1991). "Quality Ladders in the Theory of Growth". *Review of Economic Studies* 58.1, pp. 43–61.

Hall, **Bronwyn**, **Manuel Trajtenberg**, and **Adam B. Jaffe** (2001). *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** (2018). "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach". *Quarterly Journal of Economics* 133.2, pp. 801–870.

Herkenhoff, **Kyle** et al. (2024). "Production and Learning in Teams". *Econometrica* 92.2, pp. 467–504.

Hill, **Ryan** et al. (2022). "Adaptability and the Pivot Penalty in Science". *Unpublished manuscript*.

Jaffe, **Adam B** (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value". *American Economic Review* 76.5, pp. 984–1001.

Jaravel, **Xavier**, **Neviana Petkova**, and **Alex Bell** (2018). "Team-Specific Capital and Innovation". *The American Economic Review* 108.4-5, pp. 1034–1073.

Jones, **Benjamin** (2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *The Review of Economic Studies* 6 (1), pp. 283–317.

Kahane, **Leo**, **Neil Longley**, and **Robert Simmons** (2013). "The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and Language Diversity in the National Hockey League". *The Review of Economics and Statistics* 95.1, pp. 302–314.

Kaltenberg, **Mary**, **Adam Jaffe**, and **Margie E. Lachman** (2021). *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** (2021). *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. (2021). "Measuring Technological Innovation over the Long Run". *American Economic Review: Insights* 3.3, pp. 303–20.

Mann, **Katja** and **Lukas Püttmann** (May 2023). "Benign Effects of Automation: New Evidence from Patent Texts". *The Review of Economics and Statistics* 105.3, pp. 562–579.

Mortensen, **Olavur** (2017). "The Author Topic Model". *Unpublished manuscript*.

PatentsView (2024). *USPTO Patent Data for Inventors and Assignees*. United States Patent and Trademark Office (USPTO). URL: %7Bhttps://patentsview.org%7D.

Pearce, **Jeremy** (2022). "Idea Production and Team Structure". *Unpublished manuscript*.

Rosen-Zvi, **Michal** et al. (2012). "The Author-Topic Model for Authors and Documents". *CoRR* abs/1207.4169, pp. 487–494.

Sarica, **Serhad** and **Jianxi Luo** (2020). "Stopwords in Technical Language Processing". *CoRR* abs/2006.02633.

Teodoridis, **Florenta**, **Jino Lu**, and **Jeffrey L Furman** (2022). *Mapping the Knowledge Space: Exploiting Unassisted Machine Learning Tools*. Working Paper 30603. National Bureau of Economic Research.

USPTO (2024). *Cancer Moonshot Patent Data*. URL: https://www.uspto.gov/ip-policy/economic-research/research-datasets/cancer-moonshot-patent-data.

Uzzi, **Brian** et al. (2013). "Atypical Combinations and Scientific Impact". *Science* 342, pp. 468–472.

Weidmann, **Ben** and **David J. Deming** (2021). "Team Players: How Social Skills Improve Team Performance". *Econometrica* 89.6, pp. 2637–2657.

Weitzman, **Martin L.** (1998). "Recombinant Growth". *The Quarterly Journal of Economics* 113.2, pp. 331–360.

Wu, **Lingfei**, **Dashun Wang**, and **James A Evans** (2019). "Large teams develop and small teams disrupt science and technologyt". *Nature* 566, pp. 378–382.

Wuchty, **Stefan**, **Benjamin F. Jones**, and **Brian Uzzi** (2007). "The Increasing Dominance of Teams in Production of Knowledge". *Science* 316.5827, pp. 1036–1039.

Xu, **Fengli**, **Lingfei Wu**, and **James A Evans** (2013). "Flat teams drive scientific innovation". *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. Perplexity quantifies the model's uncertainty; lower perplexity indicates better generalization and model fit.

Figure S3

K MEANS CLUSTERING EXAMPLE

Notes: An example in the same K=3 knowledge space 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 from the `gensim` package option *auto*. The Y-axis presents the 5 words with the largest weight within the knowledge class to word distribution for that class. The height of the bar represents the weight on that class in the Bayesian prior.

Figure S6

Aggregate topic distribution by patent types

Notes: Plots the average knowledge class distribution by patent type. The data on Green, Automation and Cancer patents provided by PatentsView (2024), Mann and Püttmann (2023), and USPTO (2024). Again the top 5 words per class shown on the Y-axis.

Figure S7

Team Statistics

Notes: Four binned scatter plots produced with `binscatter`. Each plot is taken for the average within one year, across all teams who first patented in that year. The total number of patents is defined as in equation 8 however summing over the team field instead of a patents. Volume is defined in equation 14 as the square root of team size multiplied by the weighted average of the maximum euclidean distance between team member knowledge points, and the mean distance. I refer to this weighted average as the team breadth. The bottom two panels spilt both parts of the volume measure.

Appendix

A. Estimating the Knowledge Space: Intuition

B. Counting Objects in Knowledge Space

C. Technical Appendix: LDA

D. Robustness Tests

A1. Adding a team member instead of removing one.

E. Wordclouds

A Estimating the Knowledge Space: Intuition

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. The knowledge class and inventor parameters are learnt by iterating over patent texts and allocating inventors and topics to words.

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 Blei, Ng, et al., 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 (14).

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

$$P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{w}, \alpha, \eta, \boldsymbol{T}) = \frac{P(\boldsymbol{w}|\boldsymbol{k}, \boldsymbol{\beta}) P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta}) P(\boldsymbol{i}|\boldsymbol{T}) P(\boldsymbol{\beta}|\eta) P(\boldsymbol{\Theta}|\alpha)}{P(\boldsymbol{w}|\alpha, \eta, \boldsymbol{T})} \tag{17}$$

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[15]. Define $q(\cdot)$ as an approximation to the posterior

$$q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \lambda, \gamma, \phi) = q(\boldsymbol{\Theta}|\boldsymbol{\gamma}) q(\boldsymbol{\beta}|\lambda) q(\boldsymbol{k}, \boldsymbol{i}|\phi) \tag{18}$$

$$\approx P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{w}, \alpha, \eta, \boldsymbol{T}) \tag{19}$$

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 knowledge contained in the patent p. If a patent includes a lot of words discussing

[15] 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.

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 (15) 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}[16]. 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[17]

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

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 estimating a set of θ_i and β_k, estimates a contribution share for each team member

[16]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.

[17]For a full derivation I refer the reader to the original paper by Blei, Ng, et al., 2003

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

This paper designs the treatment model around the premature death of inventors. This provides exogenous variation in team composition, and therefore the team's position in knowledge space. To demonstrate the robustness of the results in the paper I include a set of teams which add a new inventor. This gives a weakly larger knowledge field for the team, and potentially allows them to build on more or different types of prior work.

I confirm the robustness of these results by finding that teams which add a new member, and increase the number of patents targetting a specific direction see a significant increase in the probability they patent in that direction. The reverse holds for the breakthrough patents, though the results are weaker.

The fact that the direction results are stronger than the breakthrough results is not surprising. This is an endogenous choice for the team to re-direct their patents and bring in the required knowledge. If the team brings in a green expert, their future patenting will change. While achieving a breakthrough is a much less tangible goal, you may bring in someone with radical ideas, but that doesn't ensure success.

Table A1

Treatment Team Regression Estimates: Adding an Inventor

I: Breakthrough

	Dependent variable: Pr(Breakthrough)			
$D_{\tau t} \cdot \mathbb{1}$	-0.0111**	-0.0120**	-0.0028	-0.0078
	(-2.79)	(-2.91)	(-0.96)	(-1.83)
Prior work$_{\tau_1 t}$	-0.0243***	-0.0247***	-0.0104***	-0.0102***
	(-15.08)	(-15.14)	(-8.12)	(-7.97)
Volume$_{\tau}$				1.1636*
				(2.56)
N	7331	7331	7331	7331
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Cluster FE		✓	✓	✓
Time trend			✓	✓

II: Direction

	Dependent variable: Pr(Direction)			
$D_{\tau t}$(Direction) $\cdot \mathbb{1}$	0.0079***	0.0078***	0.0097***	0.0108***
	(5.64)	(5.55)	(6.47)	(6.59)
Prior work | Direction$_{\tau_1 t}$	0.0615***	0.0616***	0.0624***	0.0623***
	(66.30)	(66.35)	(66.46)	(66.40)
Volume$_{\tau}$				-0.2327*
				(-2.04)
N	70899	70899	70899	70899
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 identifier τ is unique for each team pair (τ_1, τ_2). 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 $d(\theta_p^e, \theta_p)$. This sample replaces the treatment definition with a team adding an inventor, not losing one.

E Wordclouds

These are the fifty word clouds, one for each of the estimated knowledge classes, in addition to figure 10, here the relative size of each word in each knowledge class is visible. There is significant variation in the topics learnt. Since an LDA is an unsupervised machine learning model, a typical method of analysing the results is the human intepretability of the topics. These topics are easy to identify and distinguish. This suggests a good model fit.

Figure A4

Knowledge Class to Word Distributions: Wordclouds

16)

17)

18)

19)

20)

21)

22)

23)

24)

25)

26)

27)

28)

29)

30)

31)

32)

33)

34)

35)

Notes: For each knowledge class to word distribution this plots a wordcloud, where the word size is weighted by the corresponding probability of using that word, when discussing that class. The model does not generate names for each topic, these can be assigned by the econometrician.