Teams and Text:
Collaborative Innovation in the Knowledge Space

Joseph Emmens[‡]

November 13, 2024

Click here for the latest version.

Abstract

In this paper, I study the impact of an expanding scientific and technological frontier on team innovations. To do so, I present a novel framework that integrates inventor teams and their patent texts. I model collaboration directly through a Bayesian model of Natural Language Processing. Applied to patent text data, this model builds a map of inventors, teams, and research fields, referred to as the *knowledge space*. Trained on over 400,000 U.S. patents from the USPTO *PatentsView* database, this framework allows me to tackle unanswered questions on how teams create new knowledge. Specifically, I investigate the effect of prior work on a team's ability to produce a breakthrough—an innovation that sparks a new and successful research field. Leveraging high-dimensional patent text data, I back out two new measures: breakthrough patents and a team's knowledge field, the set of research fields accessible to the team. I combine this with data on premature inventor deaths as a quasi-natural experiment. This identifies how team innovations change as they pivot to more or less advanced research fields. The framework unifies key elements of collaboration. Teams build on existing knowledge, and prior work both supports and obstructs innovation. I show that teams generate more breakthroughs when building on enough prior work to incorporate valuable knowledge, but not so much as to stifle novelty.

Keywords: Teams, Innovation, Patents, Topic Modelling

JEL: O31 O32 O34 C55

[‡]IAE CSIC, Universidad Autònoma de Barcelona and The Barcelona School of Economics. I am deeply appreciative of the guidance and support I have received from my supervisor, Hannes Mueller, and would like to extend special thanks to Christian Fons-Rosen for their immensely valuable insights and feedback. I would also like to thank Pau Milan, Inés Macho Stadler, David Perez Castrillo, Olav Sorensen, Florenta Teodoridis, Lukas Freund and Donald Bowen III, all particpants at the WEFI, ENTER Jamboree and BSE Jamboree conferences, the Tilburg Macro Study Group and my IDEA PhD colleagues for their support and invaluable insights to improve the paper. 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. Email: joseph.emmens@bse.eu

1 Introduction

Organising inventors into effective teams is essential for growth but also for addressing society's greatest challenges. Literature suggests that the dominance of teamwork is partly driven by an ever-increasing knowledge stock (Jones, 2009), as growing fields present increasingly complex problems. However, this prior literature has largely measured innovation value through citations, overlooking how teams contribute to the creation of new and successful research fields.[1]

In this paper, I study the impact of an expanding scientific and technological frontier on team innovations. To do so, I present a novel framework that integrates inventor teams and their patent texts. I model collaboration directly through the lens of a Bayesian model of Natural Language Processing (NLP). Applied to patent text data, this model builds a map of inventors, teams, and research fields, referred to as the *knowledge space*. Leveraging high-dimensional patent text data and a tractable model of collaboration, this framework allows me to answer questions on which systematic data was missing from the literature. Specifically, I study the impact of prior work on a team's ability to produce a breakthrough—an innovation that sparks a new and successful research field. Given this, I find that teams produce more breakthroughs when building on enough prior work to incorporate valuable prior knowledge; however, not too much that it becomes hard to be novel.

The analysis in this paper proceeds in two steps. I first develop a method to characterise the latent knowledge held by inventors and their patents, disentangling the individual contribution of each team member. I train the model on 408,774 U.S. patents from 214,535 teams using the USPTO *PatentsView* database. Using the trained model, I fit an additional 2.2 million U.S. patent texts into the knowledge space. This novel space allows me to back out two new empirical measures. First, a breakthrough patent is an innovation in a research field with little prior work, which afterwards grew into a vibrant research area. Second, a team's knowledge field, defined as the set of all research fields accessible to the team. In the second stage, I combine this with data on premature team member deaths (Kaltenberg, Jaffe, and Lachman, 2021). This provides a quasi-random shock to a team's knowledge field. Through a continuous treatment model, I identify how team innovations change as they pivot to more or less advanced research fields.

I document that research fields have become increasingly crowded over time, which has meaningful consequences for whether teams achieve breakthroughs. I find that the likelihood of a patent sparking a breakthrough follows an inverted-U shape with respect to prior work.

[1]Key references studying teamwork and knowledge production through citations include Pearce, 2022; Bonhomme, 2022; Ahmadpoor and Jones, 2019, consult the literature review for a more detailed discussion.

Building on some prior work boosts innovation impact, but too much stifles novelty. This translates directly to team outcomes. For teams in advanced fields, removing a member from more established areas and shifting focus to less-explored fields increases their breakthrough potential by as much as 50%. The opposite occurs for teams in early-stage fields. Removing members who provide access to more developed areas reduces the limited prior work they can build on, lowering their innovation's impact.

I contribute to the literature by constructing a unifying framework for studying teams. To demonstrate its effectiveness, I replicate two well-known findings on team composition and breakthroughs: small teams are the most disruptive, and flat teams drive radical science (Xu, Wu, and Evans, 2013; Wu, Wang, and Evans, 2019). This framework not only consolidates existing insights but also broadens the scope of team research beyond traditional metrics. Previous studies have largely focused on measuring team value through citations and examined team composition using low-dimensional categories of inventor types. This paper addresses these limitations. By combining high-dimensional patent data with a statistical model of teamwork, I develop a new method to disentangle each team member's contribution to a patent's knowledge content. This allows me to precisely locate inventors, teams and research fields in the knowledge space. This framework can then explain which novel ideas become breakthrough research areas.

I develop and apply a two-part empirical strategy to demonstrate the results. First, I represent the latent knowledge held by inventors and patent texts. I model a patent as a combination of knowledge classes. Each class represents a specific domain of expertise. For example, a car includes knowledge on engines, wheels, fuel, etc., and some on computing. The first self-driving car then increased the amount of computing knowledge in order to automate driving. The first patent to do so was novel, but what can explain why this became a breakthrough research area? I define the knowledge space as a probability simplex across a set of knowledge classes. Inventors are characterised by their position in this space. Teams innovate by combining the knowledge profile of each member. A team's knowledge field is then defined by all possible combinations of its members. This corresponds to the set of research fields available to the team. Through a simplex, this approach naturally incorporates a spatial concept by embedding a notion of distance. I can then measure the development stage of each research field available to the team by counting the amount of prior work in each area of the knowledge space. I show in this paper that the quantity of prior work a team builds on indeed explains which novel ideas become breakthroughs.

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, and Graff Zivin, 2019; Azoulay, Graff Zivin, and

Wang, 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 is determined by the prior work in this new set of research fields. I start by showing that the premature death of a team member leads to changes in the knowledge content of a team's innovation, as revealed by the language in the patent text.

The impact of a death on a team's research 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 knowledge field after a premature death. I predict how this change determines a team's ability to achieve a breakthrough. Following the death of a team member, the average treatment increases the likelihood of producing a breakthrough by 21.27% relative to the baseline.[2] However, this result hides significant heterogeneity. When I split the sample over four quartiles of the quantity of prior work in a team's knowledge field, prior to the premature death, I find important heterogeneity in the treatment effect. For teams building on advanced areas, reducing the quantity of prior work by the average treatment increases their chances of a breakthrough by 49.7%. However, for those already working in early-stage research fields, the same change reduces their chances of a breakthrough by 61.4%.

The results can be understood through the lens of an endogenous growth paradigm. Prior work exerts opposing effects on breakthroughs. On the one hand, prior work lowers the cost of innovating by providing a solid foundation. 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 idea 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 find (Bloom et al., 2020).

These findings provide guidance for policymakers. Research funding should be distributed across fields, as concentrating it in one area may obstruct breakthroughs. Diversifying funding across new and advanced fields will help teams combine ideas in novel ways and foster breakthroughs. In addition they promote the use of cross-field collaboration. For teams working on advanced fields, by searching for new team members in up-and-coming fields they can find the novelty they need to spark a new and successful field.

[2]This number is calculated using the predicted values from the regression model. The average number of patents lost from the death of an inventor is 174.22, the baseline probability of a breakthrough is 0.44, given the coefficient 0.0022.

On a technical level, this paper makes a 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 profile. 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 engineering, 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 I find 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 research 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.[3] This evidence reflects the paper's central premise: breakthrough innovations arise from 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, Jones, and Uzzi, 2007). A range of reduced form papers have looked to describe team composition and its role in explaining innovation outcomes (Uzzi et al., 2013; Xu, Wu, and Evans, 2013; Wu, Wang, and Evans, 2019). I contribute to this literature a unifying framework for teamwork that replicates a selection of these results in one model. I do so by extending the concept of building a map of innovation, developed in Fleming and Sorenson (2004), to the inventor, team and patent level. I do so by making use of the high-dimensional patent text data. In doing so I build on

[3]Using the data kindly provided online by Arts, Hou, and Gomez, 2021. They provide a dataset which identifies new words, and new n-grams in patents and which of these are later re-used. This allows them to capture both novelty and impact.

the forthcoming study from Teodoridis, Lu, and Furman (2022). They develop a Hierarchical Dirichlet Process 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.

There is a growing literature using individual wage data to explain productivity differentials and complementarities between team members' knowledge and skills (Boerma, Tsyvinski, and Zimin, 2021; Freund, 2022; Herkenhoff et al., 2024). Closest to this paper, Pearce (2022) uses technology classifications and citations to study changes to the team knowledge production function over time. However, this literature has largely been limited to studying innovation value due to a lack of models and data capable of capturing the creation of new research fields. This paper proposes an alternative knowledge production process. As the first paper to take a model of collaboration to data on patent texts, I extend this literature to study how teams produce breakthrough innovations.

Given the rising importance of teamwork, developing empirical methods to decode how teams combine individual profiles is key to understanding their production process (Ahmadpoor and Jones, 2019). I contribute specifically to the empirical literature which looks to disentangle individual contributions to team projects (Mindruta et al., 2024; Bonhomme, 2022). There is a small but important literature using highly specific case studies in which individual inputs are observed (Kahane, Longley, and Simmons, 2013; Devereux, 2018; Weidmann and Deming, 2021). The method proposed in this paper extends this to disentangling the knowledge contribution of each member. Within this literature, recent progress has been made in identifying team leaders, as part of the division of scientific labour (Wu, Esposito, and Evans, 2024; Haeussler and Sauermann, 2020). Xu, Wu, and Evans (2013) develop a highly flexible method of classifying the leaders within teams, trained on contribution-ship data. Akcigit et al. (2018) connect this to theory on how inventors build human capital and innovate through interactions with high quality team leaders and collaborators.

Finally, the third literature I contribute to is the use of natural language processing models to identify breakthrough innovations. Arts, Hou, and Gomez (2021) developed the literature beyond using citation histories by identifying the new words created by patents to measure novelty. They then capture which of these are re-used by future innovations to measure impact. Kelly et al. (2021) develop a method of identifying breakthroughs by comparing the similarity of patent texts to patents which came before and after. The concept of breakthrough in this paper builds directly on their foundation. The key contribution of this paper is to extend this to the team level to connect the novelty and impact of their innovations to the development stages of their research fields. This paper employs a two-stage approach to back out the required latent variables from text. There is a recent literature on inference concerns when using two-stage methods (Bandiera et al., 2020; Battaglia et al.,

2024). However, as discussed in the original paper, patent texts are highly dimensional and these concerns are reduced in this context.

Paper Outline The rest of the paper is structured as follows: Section 2 defines the theoretical framework; Section 3 outlines the process of inferring the knowledge space from text; Section 4 describes the empirical reduced-form strategy; Section 5 provides descriptive statistics and validation tests; Section 6 presents the main results; and Section 7 concludes. A detailed explanation of the LDA method used can be found in the technical appendix.

2 A Framework for Team Innovation

Define $\mathcal{K}$ as a set of K knowledge classes.[4] Each class represents a specialised area of understanding. Inventors innovate 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. Formally, 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

[4] 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.

for a Dirichlet distribution is the set of K-dimensional vectors $\boldsymbol{x}$ where each $x_k \in [0,1]$ and $\sum_{k=1}^{K} x_k = 1$. The value of the Dirichlet distribution is that each element in the support of a Dirichlet distribution can be treated as a K-dimensional discrete probability distribution.[5]

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.[6] 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.[7] 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. This time with a uniform prior $\alpha = \mathbf{1}$. Drawn 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 $N_p \sim G(\cdot)$.[8] Then for each word $n_{ip} = 1, \ldots, N_p$ the team draws an inventor $i \in \tau \sim \omega_p$ and from that inventor's knowledge distribution draws a class $k \in \mathcal{K} \sim \theta_i$. Given the corresponding knowledge class to word distribution, the inventor draws 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:

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

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

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

[7] 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.

[8] 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.

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 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

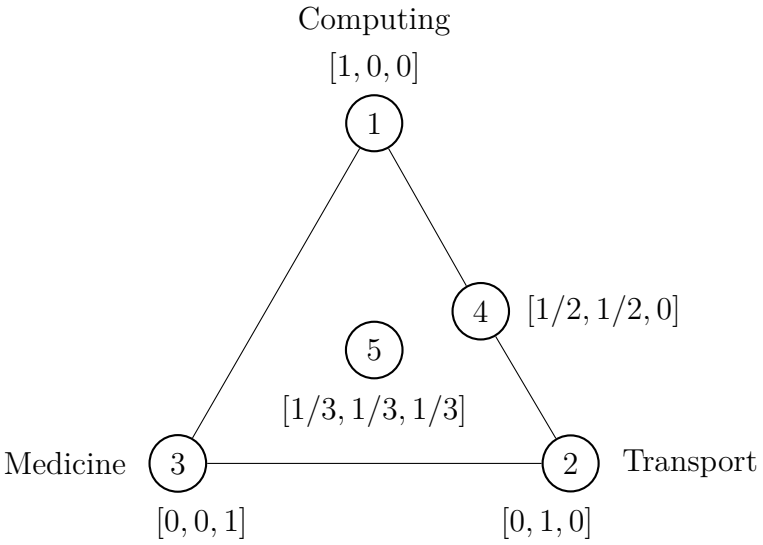
$$B(\theta, r) = \{\theta' \in \Delta(\mathcal{K}) \mid \|\theta' - \theta\| \leq r\}. \tag{3}$$

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

I define $\tilde{S}(\tau)$ 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 $\theta_p^e \in \tilde{S}(\tau)$. Formally I define the team span as the convex hull across team member distributions:

Figure 1

THE KNOWLEDGE SPACE



Notes: Example of a 2 dimensional knowledge space over 3 knowledge classes. Each point 1-5 represents either an inventor or patent knowledge profile, since both are characterised in the same space. 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.

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

To define the local knowledge field for a team consider the Minkowski sum of $\tilde{S}(\tau)$ 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 sum expands the team span into the full K-dimensions of the knowledge space. The team knowledge field is in fact the full set of patent research fields in which they will patent in expectation.

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

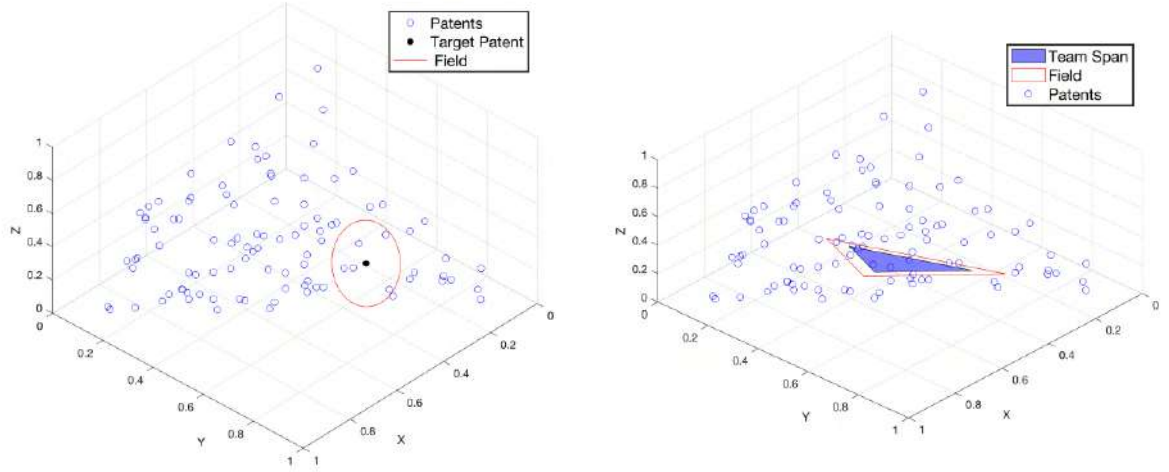
Continuing with the example outlined previously, Figure 2 demonstrates how inventors, teams and patents lie in one consistent space. Panel (A) shows an example of a patent's local knowledge field. The plot is fixed at the year patent p (shown in black) was published and there were five examples of prior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

their local knowledge field $S(\tau)$.

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. The left panel shows the research field for a target patent, shaded in black. The right panel shows the knowledge field for a team of three inventors.

2.1 Characterising Patent and Team Fields

This method backs out a latent representation of both a patent's research field, and a team's knowledge field: the set of all patent research fields on which they work in expectation. Once learnt from data, the econometrician can apply any function they desire to these objects in order to describe them and explain their role in innovation.

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 at θ_p.[9]

$$n_{pt} = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in B(\theta_p; r)\Big) \tag{6}$$

[9]A detailed explanation of how I count these objects empirically is provided in Appendix D. In short, I first slice the data by the maximum distance within the team field. I then check for the remaining patents which belong to the team field by checking the distance from each patent θ_q and the exterior of the team field.

$\mathbb{1}$ denotes an indicator function which is equal to one when the condition in parentheses is met. 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.[10]

$$b_p = \left(\frac{\text{post-count}_p}{1 + \text{prior-count}_p + \text{post-count}_p} \right) \times 100 \tag{7}$$

For a patent produced in t, prior-count$_p$ aggregates each n_{ps} for $s \leq t$ and post-count for all patents produced in $s > t$. 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 A1 gives an example that also demonstrates that the curve b_p with respect to post-count$_p$ flattens as the prior-count$_p$ increases.

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. I use these residualised values for the following reason. The measure presented in equation 7 is the raw breakthrough measure, however as made clear in Hall, Trajtenberg, and Jaffe (2001), when working with patent outcomes it is important to control for the fact that they are right-coded in time. Patents produced recently have not had enough time to be revealed as breakthroughs, since the patents that build on them have not yet arrived.

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 potentially easier to track and visualise over time. One key contribution of this paper is to extend this to the teams which produce these patents.

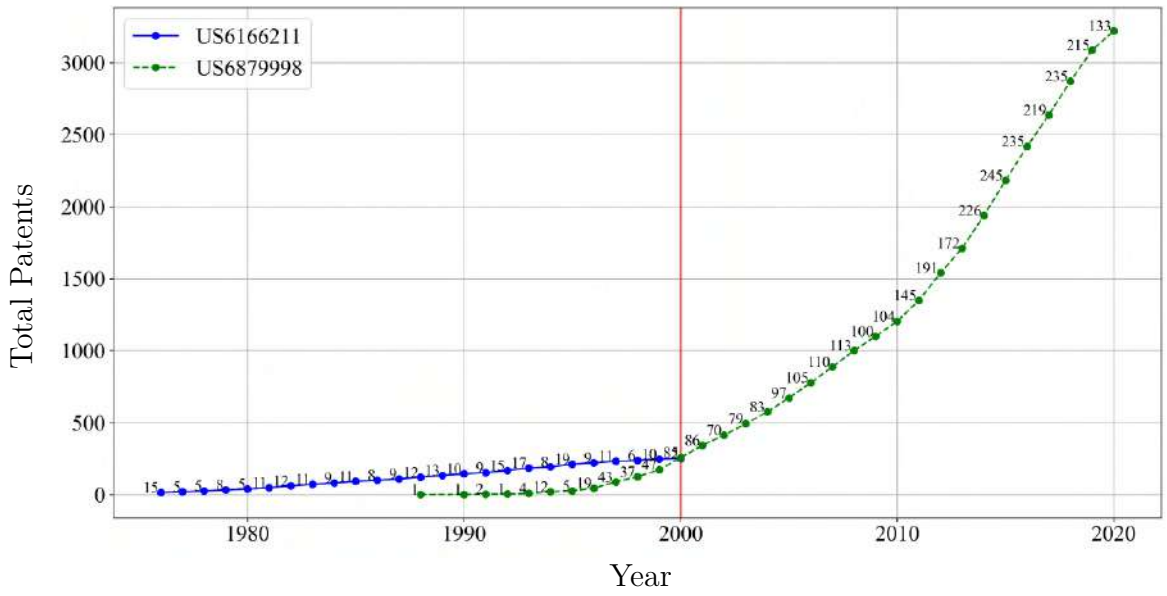
Figure 3 provides two examples of patents, one classified as a breakthrough and the other not. Both patents were applied for in the same year, however in different locations in the knowledge space. Then the two research fields then developed along very different paths over time. The Y-axis plots the total number of patents within each patent's local research field. Patent *US6879998* titled *Viewer Object Proxy* scores very highly. They patented in a space with a some prior work, 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 further patents came after them.

At the team level you can define the synonymous count. The team knowledge field

[10] In section 5, I compare the patents identified as breakthroughs by equation 7 to the literature to demonstrate the precision of this method.

Figure 3

EVOLUTION OF LOCAL KNOWLEDGE FIELDS



Notes: Two cases of low and high breakthrough patents, using the estimated knowledge space. Patent US6879998 is a breakthrough patent, where patent US6166211 is not. This is the raw count, and doesn't remove year fixed effects. The vertical line identifies the publication year of both patents. The Y-axis records the total number of patents in each target patent's research field. The number above each point counts the number of patents which arrived to that research field in each year.

represents the set of patent research fields in which the team will patent in expectation. The sum in equation 6 essentially counts the quantity of unique prior work which exists in each of these knowledge fields. Unique in that it doesn't double count for overlapping fields.

$$n_{\tau t} = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in S(\tau; r)\Big) \tag{8}$$

In addition measure the volume of their field. I approximate the volume of a team local knowledge field using the following equation[11]

$$\text{Volume}\Big(S(\tau)\Big) = v_\tau = \sqrt{m} \times (\mu \cdot D_{\max} + (1-\mu) \cdot D_{\text{mean}})) \tag{9}$$

Where m denotes team size and $\mu \in [0, 1]$.[12] Here $D_{\max}$ 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.

[11]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.

[12]In the final model I set μ equal to 0.7, to emphasise the total breadth within the team span. However, I have run the model for many alternative levels and the results don't change.

2.2 Testable Predictions

Using this framework I derive a set of testable hypotheses. In appendix section B, I present a more rigorous derivation of these hypotheses which arise directly from the innovation production process.

As a research field develops, it moves up the ladder of development, and this represents an increase in the quality of existing knowledge. The cost of producing follow-on innovations is then decreased as inventors build endogenously on the work that came before them (Grossman and Helpman, 1991). However, faced with an increasing knowledge stock, inventors suffer from the "burden of knowledge". As the innovation frontier expands, this represents an increased cost to inventors of reaching and developing on this frontier (Jones, 2009). As a team's local knowledge field populates with patents, it becomes harder to produce a truly innovative idea (Bloom et al., 2020). The compound definition of a breakthrough as both novel and impactful leads me to the following hypothesis at the patent research field level.

Hypothesis 1. *There is an inverted-U shaped relationship between the quantity of prior work in a patent's research field and the likelihood of the patent being a breakthrough.*

This implies that as the number of existing patents within a research field increases, the probability that a new patent is a breakthrough first increases due to knowledge accumulation, then decreases after a certain point due to saturation.

In that case, how can a team produce a new breakthrough idea? For a team composed of inventors covering a well-established research area, finding a novel idea is challenging. Increasing the novelty of their work might require removing a member contributing knowledge from the most developed fields. However, for a team in an under-explored area, this same adjustment would weaken their chances. It would strip away the limited knowledge they possess, setting them further back on the ladder of development. I propose to test this with an additional hypothesis at the team level. Importantly, when teams remove (or add) a member they also change the size of the team's field, in doing so they change the quantity of potential combinations. The effect of building on a few patents dispersed across a vast set of possible combinations is likely different from building on the same number within a smaller, more concentrated field. The following hypothesis includes this feature by using variation in the density of prior work within a team's field.

Hypothesis 2. *The impact of reducing the density of patents in a team's knowledge field on the probability of their next patent being a breakthrough depends on the development stage of their initial field:*

- *If the team spans an advanced research area, moving to areas with a lower density of prior work increases their breakthrough probability.*

- *If the team spans an early-stage research area, moving to areas with a lower density of prior work decreases their breakthrough probability.*

The rationale for why is as follows. Prior work enhances the impact of an innovation; thus, when a team incorporates more related work, the quality of their innovations improves. However, 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.

3 Inferring 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 of prior work within a team's local knowledge field as to test both hypotheses.

3.1 Data and Sample

I build the knowledge space from US patent data from *patentsview*, the online data base for the United States Patent and Trademark Office (USPTO). I restrict the sample to teams who applied for their first patent after 1990, and their last prior to 2011. I build the sample around three types of teams, which I combine into a panel of team, patent observations.

The first team type 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, Jaffe, and Lachman (2021). I define a premature death using the following logic. I take one unique death date per inventor[13], 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 from the sample. Given

[13]This data set was produce by scraping four well known US public record databases, for many inventors they scraped multiple potential birth and death dates. They score each one according to their belief that it is an accurate measure. I take the maximum observation with a maximum score. For more details see the original paper.

the delay in producing a patent, returning in less than five years is relatively fast to turn around a new patent. I claim that the death was a quasi-natural experiment in changing team composition, I discuss this strategy in more detail in section 4.2.

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 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. The second provide an endogenous team composition change that allows me to study adding new members, as a robustness check. 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. However, since I am using a conditional logit model, I estimate the treatment effect on teams which switch outcomes at least once. In other words they produce least one breakthrough. The final sample includes the following split: 72 teams which don't replace the prematurely deceased inventor, 510 which do replace them and 1709 control teams.

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 of 408,774 patents written by 270,065 inventors. This sample contains patents and inventors for which I don't track their entire history, however they help provide a precise measure of knowledge for the target 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.[14] 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 trained model, I fit each patent into the estimated knowledge space.

I combine additional data for the robustness check, and additional section demonstrating the knowledge space. Firstly 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).

[14]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

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 (Cancer Moonshot: 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.*

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, and Jordan, 2003 and have become a popular method of NLP. Consider this a brief and intuitive overview of how an LDA infers a set of parameters which approximate the knowledge space. For a full description consult the accompanying technical description in appendix section E.

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: K is the number of knowledge classes. η the Bayesian Dirichlet prior on the knowledge class to word distribution. Iterations sets the number of cycles used to update the knowledge class distributions, passes are full the number of times the model goes over the entire dataset, and the gamma threshold sets the stopping point when the difference between topic updates is sufficiently small. The model has been run various times changing these parameters, and the results remain 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, McMahon, and Prat (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.[15]

[15]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.

Figure A2 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.

The words contained in a patent describe its design and use. The LDA model reduces the dimension from over 250,000 words in the raw patent texts 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 G, 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.

Figure A4 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.

4 Empirical Strategy

I present a set of regression models to test both hypotheses derived in section 2. 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 probability that patent produces a breakthrough.

4.1 Hypothesis 1: Patent Level

Here the dependent variable varies at the patent level, where each patent maps into one team τ and application year t. The regression is run as as a standard logit model to predict whether patent p from team τ in year t is a breakthrough, or not. The full specification is given by

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

where

$$X'_{\tau t(p)}\boldsymbol{\psi} = \beta_0 + \delta_t + \beta_1 n_{pt} + \beta_1 n_{pt}^2 + \beta_2 d_p + \beta_3 m_\tau \tag{11}$$

The main parameters of interest are β_1 and β_2. $\beta_1 > 0$ and $\beta_2 < 0$ are consistent with an inverted-U shape. 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 include year fixed effects for multiple reasons. They control for the fact that breakthroughs are right-coded in time: patents published recently have not yet had chance to be realised as breakthroughs.

4.2 Identification strategy for Team Outcomes

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. I define premature as having died within three years of patenting. The number three is chosen as in the USPTO raw data, teams on average patent every three years. Therefore if an inventor dies within three, it is reasonable to assume that on average this would change their next patent outcome. This definition is therefore based around them being active, not age or health status, and is in line with the literature Azoulay, Fons-Rosen, and Graff Zivin, 2019. Denote the initial team, prior to the premature death as τ_1. This team must return to patent within 5 years denoted as τ_2 to entire the sample. This is to define a cap on the number of years in which they must return. 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.[16] The identifying

[16]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.

assumption here is that the death is an unexpected event, where the impact on the team's knowledge field is captured by $D_{\tau t}$.

Notice that the treatment is time dependent. If a team member prematurely dies in period t then $D_{\tau t}$ measures the contemporaneous change in the quantity of prior on which the team builds. However for all future periods this variable captures the knowledge foregone by the untimely death. The idea being that if the inventor had not passed away, the team could have continued to patent in those research fields. 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.

4.3 Hypothesis 2: Team Level

This is model is run on a team patent panel. Each patent is a new period s, such that the team τ is repeated over their 1^{st}, 2^{nd}, 3^{rd} patents and so on. I then predict the probability that $Y_{\tau s}$, a team's nth patent, is a breakthrough.

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

The full set of independent variables is given by,

$$X'_{\tau s}\boldsymbol{\psi} = \alpha_\tau + \mu_s + \delta_t + \beta_1 n_{\tau_1 s} + \beta_2 D_{\tau s} + \beta_3 v_\tau + Z'_p\boldsymbol{\delta} \tag{13}$$

Hypothesis 2 requires that the density of patents changes within the team local knowledge field. Therefore I introduce a control for the volume denoted v_τ, as defined in equation 9. $n_{\tau_1 s}$ controls for the quantity of prior work within the team field of the initial team, prior to the inventors premature death. β_2 then captures the treatment effect of removing patents from the team's knowledge field. In other words, the effect of moving them into a less explored area of the knowledge space.

To test the compound hypothesis, I split the sample of teams into quartiles of prior work $n_{\tau_1 s}$. I then run the same regression as specified in equation 12 for each quartile separately. For those teams initially building on a lot of prior work ($n_{\tau_1 s}$ high), $\beta_2 > 0$ is consistent with the gain from them moving into under-explored areas and drawing more novel ideas. Conversely, for those teams initially building on a little prior work ($n_{\tau_1 s}$ low), $\beta_2 < 0$ is consistent with them losing out by having fewer prior examples to incorporate, reducing the impact of their work.

5 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 shows two sets of descriptive statistics. The first panel describes the sample used to train the LDA model. The second panel is a sub-sample of the first, and describes the teams and patents used in the reduced-form regression model.

Table 2
DESCRIPTIVE STATISTICS

LDA sample				
Patents	Obs	Mean	Min	Max
Team size	408774	3.423	1	76
% Breakthrough	408774	0.260	0	1
Specialisation	408774	0.455	0.150	0.975
Hierarchy	408774	0.047	0	0.928
Inventors	Obs	Mean	Min	Max
Teams	270,065	3.078	1	767
Patents	270,065	5.181	1	4549
Specialisation	270,065	0.533	0.028	1
Contribution weight	270,065	0.263	0*	1
Treatment sample		1990-2010		
	No Replace	Replace	Control	
Treatment Status	72	510	1709	
	Obs	Mean	Min	Max
Team size	10,934	2.470	1	20
Team patents	10,934	7.399	2	51
% Breakthrough	10,934	0.449	0	1
Total Count	10,934	185.565	0	3198
Volume	10,934	0.637	0	4.69
Density	7,929	329.03	0	35327.19

Notes: Volume is defined by the square root of team size, multiplied by the weighted average of the maximum and mean distance between team member knowledge profiles, as given in equation 9. Total count is defined as the number of patents within a team's or patent's knowledge field, given by equations 6 and 8 respectively. Density is defined as total count divided by the volume, and is set to missing as some teams of size 1 have zero volume and zero count. * since this is approximately zero in the data. The treatment sample split is conditional on them being part of the final conditional logit sample- they have at least one breakthrough patent.

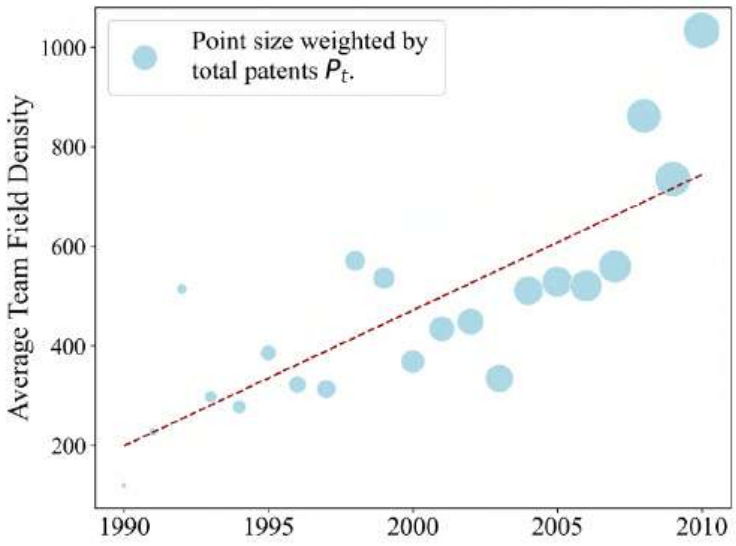
5.1 Aggregate Statistics

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 area on which a team works develops over time.

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 4 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 increases. This doesn't however mean that the number of patents within a team's local knowledge fields increases mechanically. 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.

Figure 4
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. Volume is defined by the square root of team size, multiplied by the weighted average of the maximum and mean distance between team member knowledge profiles, as given in equation 9. 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 in the knowledge space in each year.

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. For the volume to remain constant then, teams must be combining inventors who are closer together, such that the maximum and mean distance between members is decreasing. I show that this is in fact the case and the comparison across team statistics and the breakdown of the volume measure can be seen in Figure A6.

5.2 Breakthrough Patents

This paper presents a novel empirical concept for breakthrough research fields. Table 3 provides a set of validation statistics to demonstrate the empirical power of the framework.

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)	$\%\Delta$ new re-used words in breakthrough patents	8.67^{***}
	$\%\Delta$ new re-used bi-grams in breakthrough patents	47.6^{***}
	$\%\Delta$ new re-used tri-grams in breakthrough patents	44.1^{***}
Citations	$\%\Delta$ forward citations for + $\Delta 1\%$ in post count$_p$	$2.07\%^{***}$
	$\%\Delta$ backward citations for + $\Delta 1\%$ in prior count$_p$	$1.09\%^{***}$
	$\Delta\%$ forward citations for breakthrough patents	$16.2\%^{***}$

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

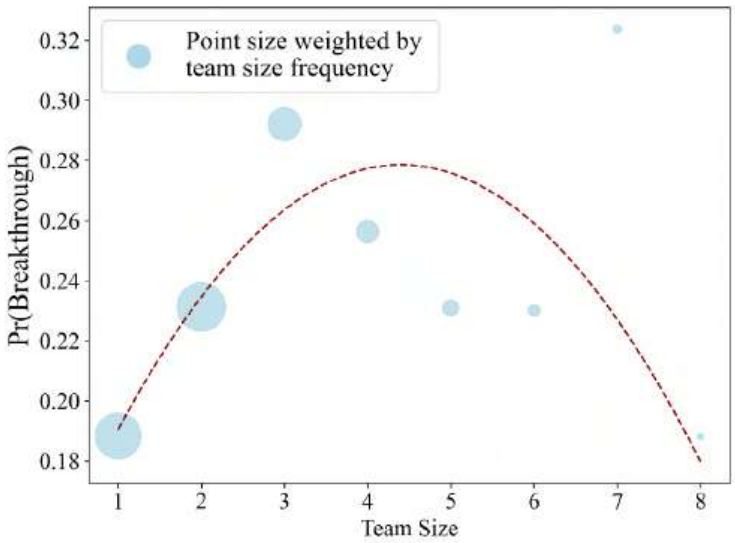
This paper develops on the work in Kelly et al. (2021) and using their data I find the correlation between their binary breakthrough classification and the one produced in this paper. I find a positive correlation of 0.221. I extend this and show that their continuous breakthrough score is negatively correlated with the prior-count of patents belonging to that local knowledge field, but positively correlated with the post-count.

In addition, using the Arts, Hou, and Gomez (2021) data I first show that patents which I classify as breakthrough patents contribute 8.57% 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, 47.6% new word pairs, and 44.1% new-three word tuples. This result speaks to the central premise on

how innovation occurs, through recombining existing knowledge. This data is particularly useful in that it allows for the comparison across simply being novel, and being novel and later reused. This is because the data presents two counts for each patent, the number of new n-grams, and the number of these n-grams what were later reused.

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 2.07% 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 1.09% more backward citations.

Figure 5
BREAKTHROUGH INNOVATIONS AND TEAM SIZE



Notes: The Y-axis plots the percentage of patents classified as breakthroughs, produced by teams of each discrete team size from 1-8. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Point size is weighted by the frequency of team sizes, since they are discrete bins and not equally sized.

Having validated this measure I replicate the first result from the literature in this unifying framework for teams. Wu, Wang, and Evans (2019) show that small teams disrupt science, while large teams develop it.[17] Figure 5 plots the probability of producing a breakthrough by team size. I plot up to a team size of 8 as this corresponds to 99% of the data.

[17]This paper uses a measure of innovation disruption. Disruption is measured by examining the citation patterns of future papers that reference a given paper. Specifically, they calculate a "disruption score" that reflects the extent to which a paper makes prior work obsolete or shifts the research direction.

The graph confirms that teams outperform working alone, as all team sizes above 1 outperform solo patents. Teams of 3 work best, and team size is negatively correlated with breakthroughs beyond that. However the inverted-U shape is also foretelling the results on the role of prior work.

5.3 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.

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 of the 50 runs of a random forest I calculate the total feature importance for the lead and second inventor patent histories. The features are the top five most common CPC classes used by the lead, and the second inventor. The target variable is the CPC class awarded to the patent. The final T-test is calculated over N=50.

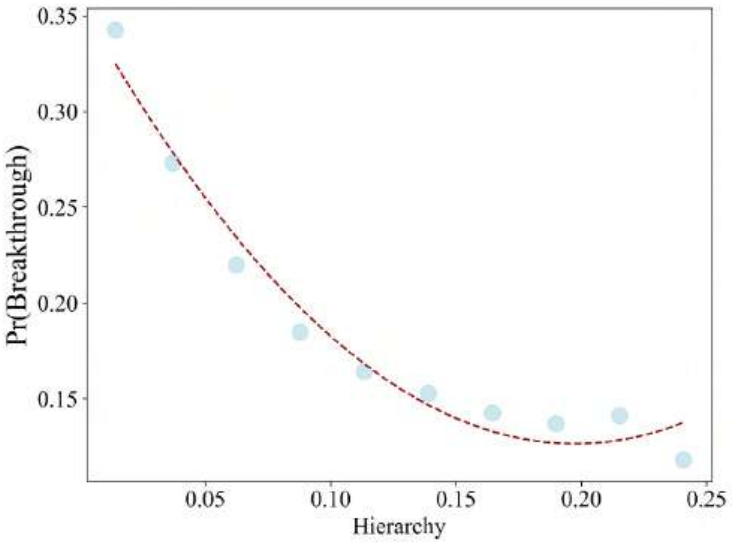
For each patent in the sample I define the lead and second inventor by ordering their estimated contribution shares and calculate the percentage difference between them. With a random forest, I predict the CPC classifcation awarded to a patent with two sets of explanatory variables: the five most common 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 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.

Figure 6
Breakthrough Innovations and Hierarchy



Notes: Hierarchy is measured by taking the vector of contribution weights and finding the euclidean distance from a vector of length m (team size) in which all inventor contribute $1/m$. The Y-axis plots the average breakthrough value for 10 equally sized bins of the hierarchy measure. The breakthrough classification is based on Equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent).

Having validated this measure I replicate a second well known result from the literature on team composition and breakthrough innovations. Xu, Wu, and Evans (2013) show that hierarchical teams produce fewer breakthroughs that teams in which members contribute more equally. I replicate this result by taking the vector of contribution weights ω_p and finding the euclidean distance from the vector of length m in where all inventor contribute

$1/m$. This measure is increasing in the hierarchy of the team, and is minimised at 0 when all team members contribute equally. Clearly from Figure 6, the result replicates, and more equally distributed teams produce more breakthroughs.

6 Main Results

I first present a set of results that demonstrate how team innovations change in response to them pivoting to new research fields. This first sub-section can be skipped for those readers interested in the main breakthrough results. I then present the main results on how teams produce breakthrough innovations to test the hypotheses presented in section 2.

6.1 Knowledge Content of Team Innovations

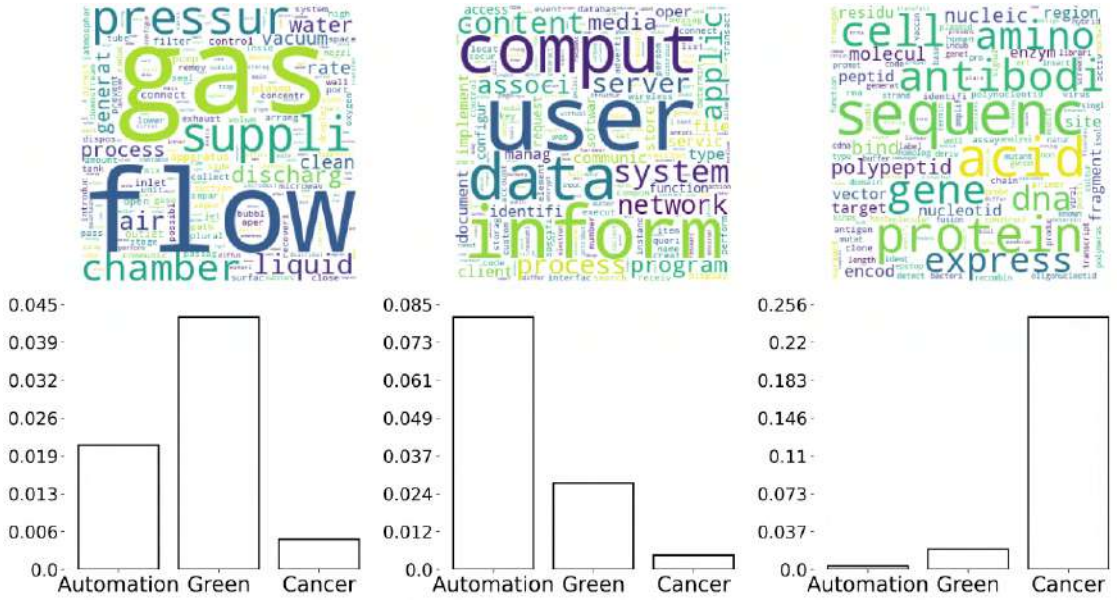
In this section, I examine how the knowledge content of team innovations shifts as teams move into new research fields. I introduce three binary classifiers for the knowledge content of each innovation, which effectively segment the high-dimensional knowledge space into binary categories. This partitioning allows us to fit these classifiers into a regression model to show how team composition influences innovation outcomes. As team members are added or removed, the team's potential knowledge combinations change, and therefore the research fields available to them. This directly affects their innovation output. This section illustrates how the knowledge content of their patents also depends on the history of prior work in each research field.

Each patent is classified by z_p where $z_p = 1$ if patent p achieves direction z. For this paper I take three exogenous classifications of whether each patent in the knowledge space achieves that purpose, or not. The directions are the following. Does the patent save labour? Does the patent improve cancer treatment? Does the patent mitigate the negative effects of climate change? These classifications are taken as exogenous (PatentsView, 2024; Mann and Püttmann, 2023; Cancer Moonshot: USPTO, 2024). Further details are presented in the data section 3.1. I combine the three directions in order to show how team innovations respond to past work, without focusing on any specific technology or field.

I examine how variation in the words used in a patent reflect the technological direction of that patent. For example, by comparing the most frequent knowledge classes across patents that mitigate climate change, target cancer treatment or produce automated technologies, we can see how each purposes is reflected in the patent vocabulary. Figure A5 shows the average weight for all knowledge classes split over three patent types. In Figure 7, I present three of the 50 estimated knowledge classes, and the average weight for patents of each type.

Figure 7

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. The word size reflects the probability of using that word when describing that knowledge class. The full set of word clouds can be found in appendix section G.

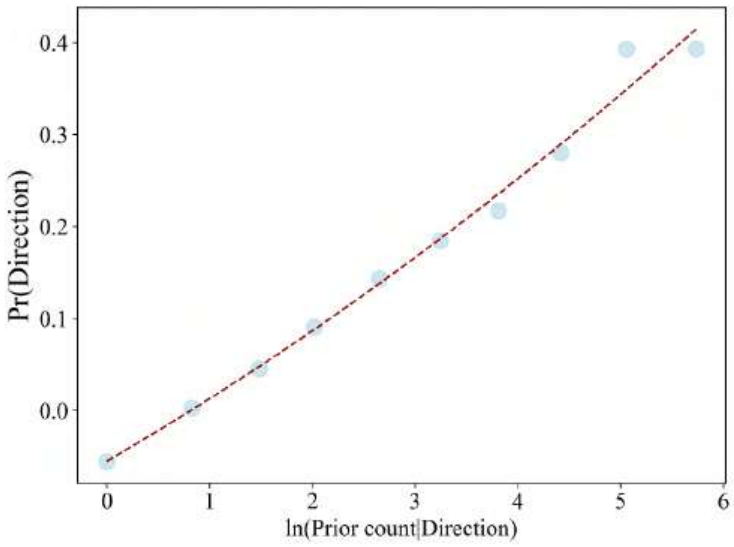
Clearly patents which target cancer treatment use can be distinguished as using words such as *cell*, *antibody*, *gene*, while automation patents use *computer* and *information*. This figure supports the empirical concept of the knowledge space, that the patent text is informative of the knowledge content of an innovation.

I first show that prior knowledge shapes future innovations. Take the patent level count of prior work defined in equation 6 and include an additional condition to the indicator function: that $z_p = 1$. This will count the quantity of prior work in that field which achieves direction z. In Figure 8, I show that the probability a patent targets direction z is increasing in the quantity of prior work in that research field which also targets z. This is demonstrated more rigorously in Table A1 where each additional patent increases the probability a patent targets direction z for between 2 to 4 percentage points.[18]

Importantly, Figure 8 shows no non-linear effect. This demonstrates two important features of endogenous growth. Prior work reduces the cost of future innovations, but also leads to path dependence. Path dependence refers to how the direction and nature of future

[18]In this table, and the later treatment model I stack the three directions into one regression model and include a period × direction fixed effect. This leads to an tripling of the sample size, and the effect is now the average across each direction. This achieves the goal of presenting technologically neutral results.

Figure 8
PATENT DIRECTION



Notes: This figure plots the probability that a patent achieves a given direction z by the log count of the number of prior patents existing in the local knowledge field of that patent, which also target z. The three directions are 1) mitigate climate change (PatentsView, 2024), 2) improve cancer treatment (Cancer Moonshot: USPTO, 2024) and 3) automate production (Mann and Püttmann, 2023). All three are stacked into one model.

innovation is determined by past work. Where early stage advances establish a path that is difficult to change. Aghion et al. (2016) show how changing the direction of a research field, for example to go green, is a challenge since the relative cost of producing either green or dirty patents is a function of what came before. This can be seen in Figure 8 as each succesive patent targeting a given direction increases the chances of future work doing so further.

At the team level, this linear effect leads to straightforward outcomes. Here I use the same treatment as defined in section 4, however again adding the new condition that the patent belongs to the team's field, and targets direction z. Therefore the treatment now captures how many prior patents the team loses following the premature death of a colleague. Again using the stacked regression model, I find that the probability a team's next patent targets a given direction z decreases by around 1 percentage point, for each prior-patent targeting the same direction removed. Naturally, following the premature death of an inventor, teams that lose access to the required knowledge to produce a patent of a certain type, see a change in the knowledge content of their innovations.

Table 5
TEAM TREATMENT ESTIMATES: DIRECTION

	Dependent variable: Pr(Direction)			
	1.	2.	3.	4.
$D_{\tau t}$ | Direction	-0.0081***	-0.0174***	-0.0097***	-0.0108***
	(-5.64)	(-7.43)	(-4.64)	(-4.97)
Prior work$_{\tau_1 t}$ | Direction	0.0114***	0.0419***	0.0217***	0.0217***
	(11.09)	(40.33)	(29.33)	(29.40)
Volume				-0.8094**
				(-2.61)
N	32802	25287	25287	25287
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period × Direction FE		✓	✓	✓
Year × Direction FE			✓	✓

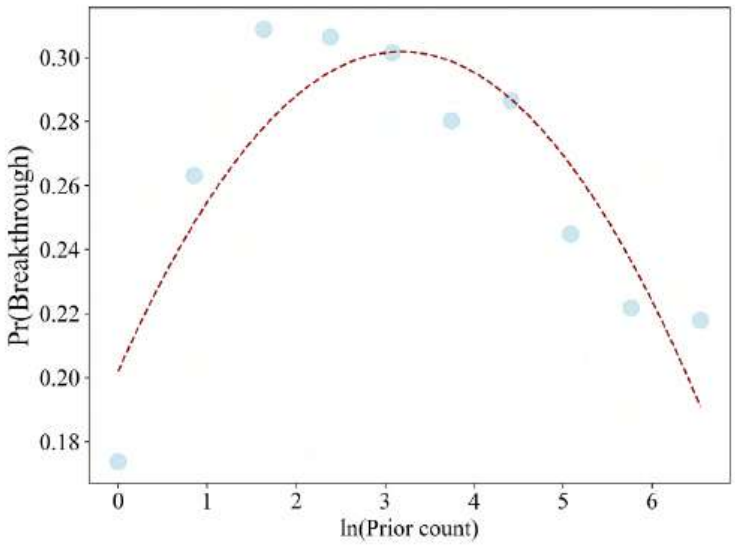
Notes: The first column uses a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each pair (τ_1, τ_2). The dependent variable is a stacked indicator for whether a patent achieves the given direction. The three directions are 1) mitigate climate change (PatentsView, 2024), 2) improve cancer treatment (Cancer Moonshot: USPTO, 2024) and 3) automate production (Mann and Püttmann, 2023). Controls include $d(\theta_p^e, \theta_p)$.

6.2 Breakthrough Innovations

I present the main results to test the hypotheses laid out in section 2 using the empirical strategy detailed in section 4.[19] I first show supporting evidence for Hypothesis 1. Figure 9 shows that the probability a patent becomes a breakthrough is an inverted-U shape in the quantity of prior work on which it builds. Recall the definition of a breakthrough patent using equation 7. Prior-count appears in the denominator, therefore a the derivative would be negative. However, we see that for low levels of prior work, this function is increasing. This is evidence that prior work increases the impact of an innovation, and in fact post-count is determined in some part by what came before. However, the function later inflects and prior work becomes a barrier for novelty. As prior work accumulates, teams find it harder to be novel and this effect wins out, thus turning the slope back to the negative coefficient expected from the breakthrough definition.

[19]I use the breakthrough measure defined endogenously by the knowledge space in equation 7. To remove concerns that this may be driven by some mechanical feature of the model I replicate all results using the Kelly et al. (2021) data in appendix section F.

Figure 9
PATENT BREAKTHROUGH



Notes: This figure plots a binned scatter plot and fitted regression line. The log count of the number of pre-existing patents in a patent's research field is split into 10 equally sized bins and the Y-axis plots the probability of a breakthrough within each bin. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent).

I test Hypothesis 2 at the team level. I first present the set of results averaging over all teams. All variables are defined as in section 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. Table 6 shows that on average, the novelty component of a breakthrough wins out, and teams benefit from moving to under explored areas. This frees them from established paradigms, as they produce more breakthrough ideas.

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 174. This change results in a 9.55 percentage point increase in the probability of producing a breakthrough, this represents a 21.28% increase on the baseline.[20] For those that replaced the inventor though,

[20]The change in probability is calculated using the baseline probability of 0.449. The coefficient of 0.0022

Table 6
Treatment Team Regression Estimates

Panel A: Breakthrough

	Dependent variable: Pr(Breakthrough)			
	1.	2.	3.	4.
$D_{\tau t}$	0.0007***	0.0066***	0.0028***	0.0022***
	(7.29)	(10.96)	(6.94)	(5.51)
Prior work$_{\tau_1 t}$	-0.0009***	-0.0357***	-0.0104***	-0.0106***
	(-12.56)	(-22.85)	(-9.19)	(-9.18)
Volume$_{\tau}$				-2.2554***
				(-7.76)
N	35991	10934	10934	10934
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

Notes: The first column uses a standard logit model. Columns 2-4 are conditional logit models with 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 is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Controls include $d(\theta_p^e, \theta_p)$.

the average treatment was to only lose 43 patents. This leads to only a 5.26% increase on the baseline. By replacing the inventor, they produce fewer breakthroughs.

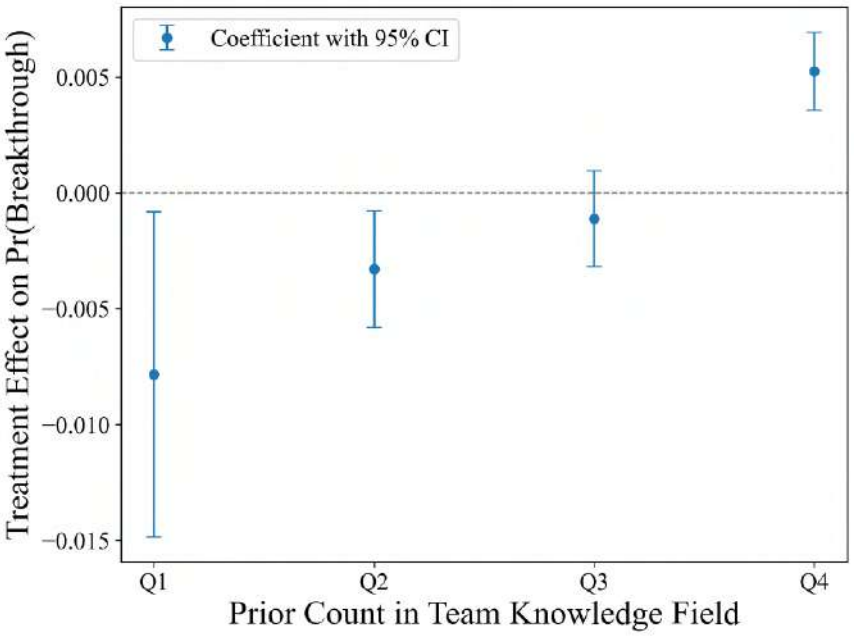
Given the inverted-U shape in Figure 9, I show how this translates into a heterogenous treatment effect. Figure 10 plots the same regression results, however split over four samples. I split the sample into quartiles of prior work in the initial teams knowledge field and run the model for each sub-sample. We see that the inverted-U shape translates directly into recommendations at the team level. For teams building on advanced areas, reducing the quantity of prior work by the average treatment of 174 patents increases their chances of a breakthrough by 49.73%. However, for those already working in early-stage research fields, the same change reduces their chances of a breakthrough by 61.4%. Importantly, the teams in early-stage areas which replace their inventor see a significantly smaller decrease in their ability to produce breakthroughs. If the number of patents lost is reduced to the average by

and an average treatment of 174.222 yield a change in log-odds of 0.3833. Applying this to the baseline log-odds of -0.204 results in new log-odds of 0.1793. Converting this back to probability gives 0.545, indicating a change of approximately 9.6 percentage points, which is a 21.4% increase relative to the baseline.

replacing the inventor, these results show a much smaller 8.1% decrease in the likelihood of a breakthrough. This demonstrates the importance of the availability of knowledge within inventor networks.[21]

Figure 10

WORDCLOUDS AND KNOWLEDGE CLASS DISTRIBUTIONS BY PATENT TYPE



Notes: This figure plots the heterogenous treatment effect for the continuous treatment variable outlined in equation 13. The x-axis plots the coefficient on the treatment for each of the four quartiles of the quantity of prior work in a team's knowledge field, prior to the premature death of a collaborator. The dependent variable is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Each model is a conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. Controls include $d(\theta_p^e, \theta_p)$.

For teams in advanced areas, removing a team member who contributes the established knowledge increases their chances of producing a breakthrough. For them, increasing the novelty of their patents is key, and therefore moving into less-explored research fields improves their ability to be novel and produce breakthroughs. However, for a team in the first quartile, those that are already building on relatively little prior work, the same change is detrimental. If they remove a member who contributes the little knowledge on which they are building,

[21]This result suggests a valuable follow-on research project which studies frictions in the *market for collaborators*. If a team suffers the premature death of a collaborator who provided a certain type of required knowledge, perhaps there is a deficit in the supply of this knowledge, and they cannot be easily replaced. This variation in post-death team outcomes may be driven by their ability to find replacement inventors.

their chances of producing a breakthrough reduce further. Their innovations may be novel, however they move too far down the ladder of development and their innovations lose impact.

7 Concluding Remarks

In this paper I ask how the development stage of a research field determines a team's ability to produce breakthrough innovations. A deeper understanding of the determinants of breakthroughs is key to modelling how the innovation frontier moves forward over time. Traditionally, the literature on knowledge production has focused on value. This paper presents a contribution to the innovation literature by constructing a unifying framework for teamwork capable of capturing the creation of new and successful research fields.

I model collaboration directly through the lens of 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. I build a mapping of inventors, teams and patents in which to study how teams innovate. I refer to this as the knowledge space. 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 model teamwork in a tractable approach. I use premature inventor deaths to identify the effect of pivoting to more or less advanced research fields on a team's ability to produce breakthrough innovations. I find a non-linear relationship between prior work and breakthroughs. I find that teams produce more breakthroughs when building on enough prior work to incorporate valuable prior knowledge, but not so much that it stifles novelty.

The framework presented here marks the beginning of a rich future research agenda. 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 we produce science and technology.

References

Aghion, **Philippe**, **Antoine Dechezleprêtre**, **David Hémous**, **Ralf Martin**, and **John Van Reenen** (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**, **Santiago Caicedo**, **Ernest Miguelez**, **Stefanie Stantcheva**, and **Valerio Sterzi** (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**, **Andrea Prat**, **Stephen Hansen**, and **Raffaella Sadun** (2020). "CEO Behavior and Firm Performance". *Journal of Political Economy* 128.4, pp. 1325–1369.

Battaglia, **Laura**, **Timothy Christensen**, **Stephen Hansen**, and **Szymon Sacher** (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.

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**, **Charles I. Jones**, **John Van Reenen**, and **Michael Webb** (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*.

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

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.

Haeussler, **Carolin** and **Henry Sauermann** (2020). "Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity". *Research Policy* 49.6, p. 103987.

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**, **Jeremy Lise**, **Guido Menzio**, and **Gordon M. Phillips** (2024). "Production and Learning in Teams". *Econometrica* 92.2, pp. 467–504.

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.

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**, **Dimitris Papanikolaou**, **Amit Seru**, and **Matt Taddy** (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.

Mindruta, **Denisa**, **Janet Bercovitz**, **Vlad Mares**, and **Maryann Feldman** (2024). "Stars in Their Constellations: Great Person or Great Team?" *Management Science*.

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**, **Thomas L. Griffiths**, **Mark Steyvers**, and **Padhraic Smyth** (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.

Uzzi, **Brian**, **Satyam Mukherjee**, **Michael Stringer**, and **Ben Jones** (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.

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

Wu, **Renli**, **Christopher Esposito**, and **James Evans** (2024). *China's Rising Leadership in Global Science*. Working Paper 2406.05917. ArXiv.

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.

Teams and Text: Collaborative Innovation in the Knowledge Space

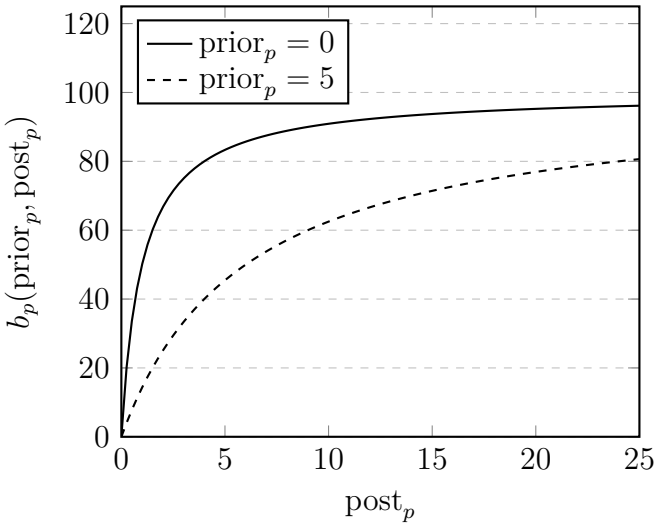
Appendix

Table of Contents:

A Additional Tables and Figures

Figure A1

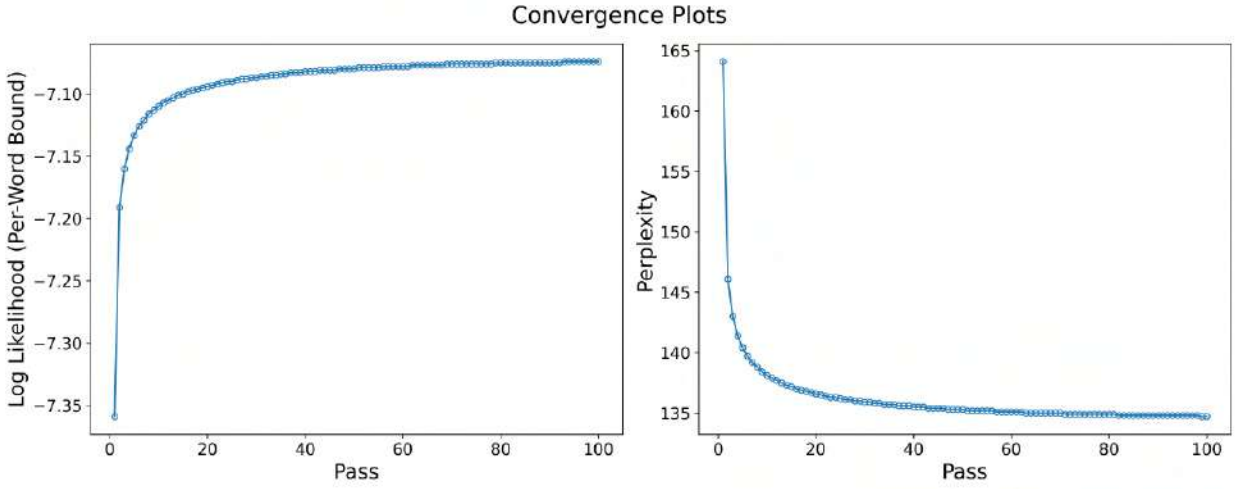
RAW BREAKTHROUGH MEASURE



Notes: Function for the raw breakthrough measure at the patent level plot over a generated range of post-count values. 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.

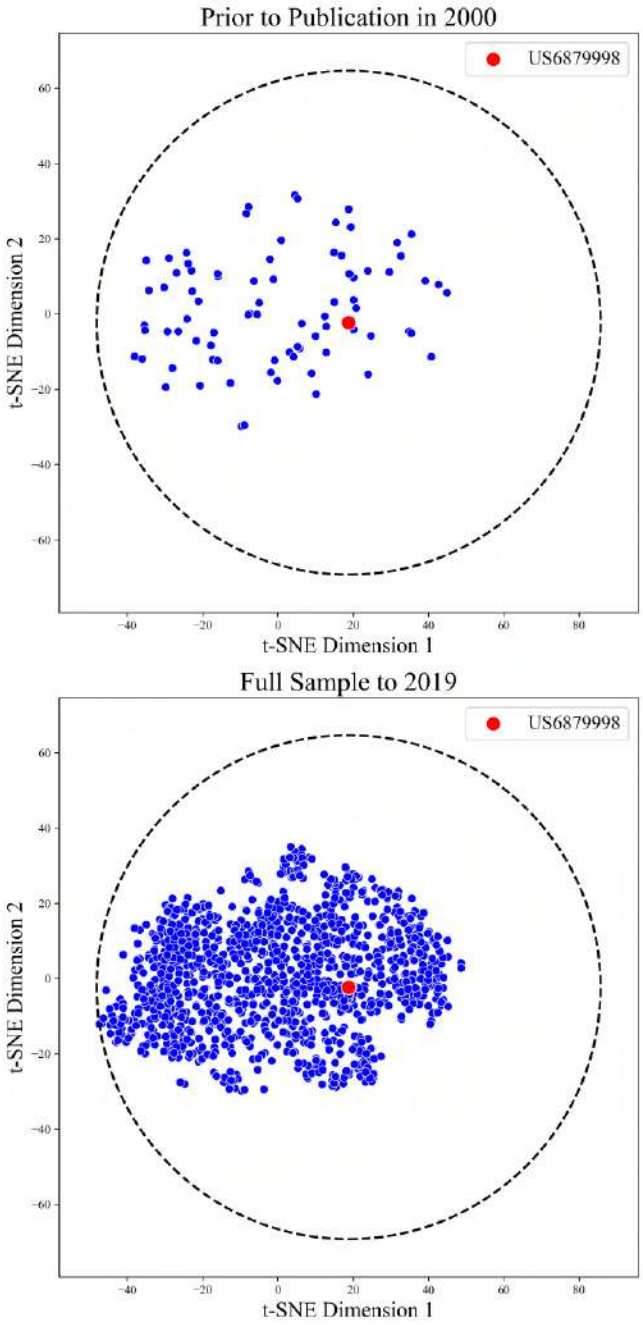
Figure A2

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.

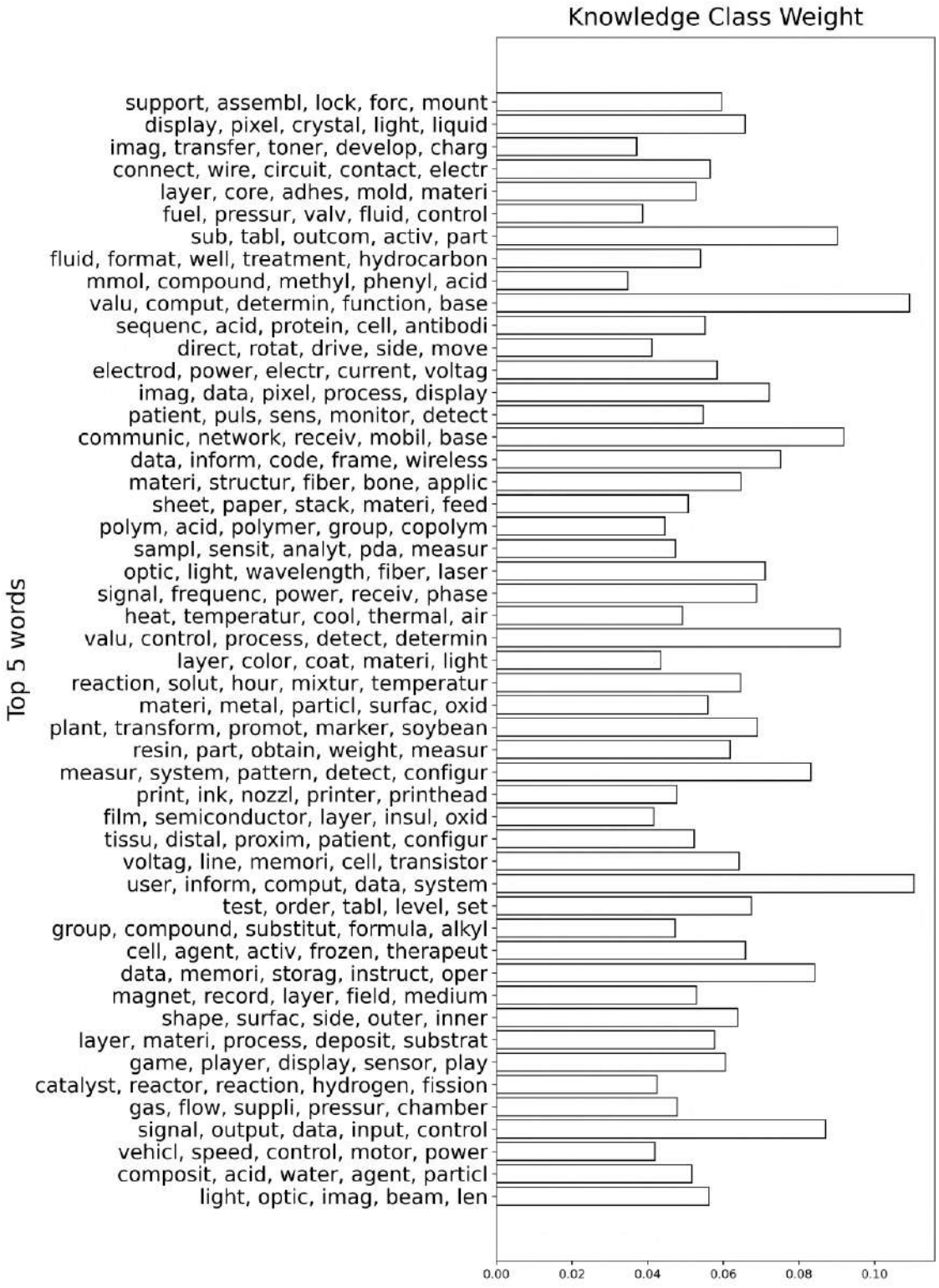
Figure A3
VISUALISING THE 50-DIMENSIONAL PATENT FIELDS



Notes: Pre & Post Publication for US6879998 titled " Viewer Object Proxy" from Google Inc. This figure visualizes the t-SNE embeddings of patent topic distributions for a selected target patent, US6879998. The 50-dimensional patent embedding θ_p is reduced to two dimensions using t-SNE, a dimensionality reduction technique optimized for capturing relative similarities between points in lower-dimensional space. In each panel, the red marker highlights the target patent, while other blue markers represent additional patents. The top panel shows only patents published prior to or in the same year as the target patent, while the bottom panel includes the full sample. A dashed black circle, centered on the target patent, encompasses the maximum distance to other patents in the sample.

Figure A4

INFERRED BAYESIAN PRIOR α



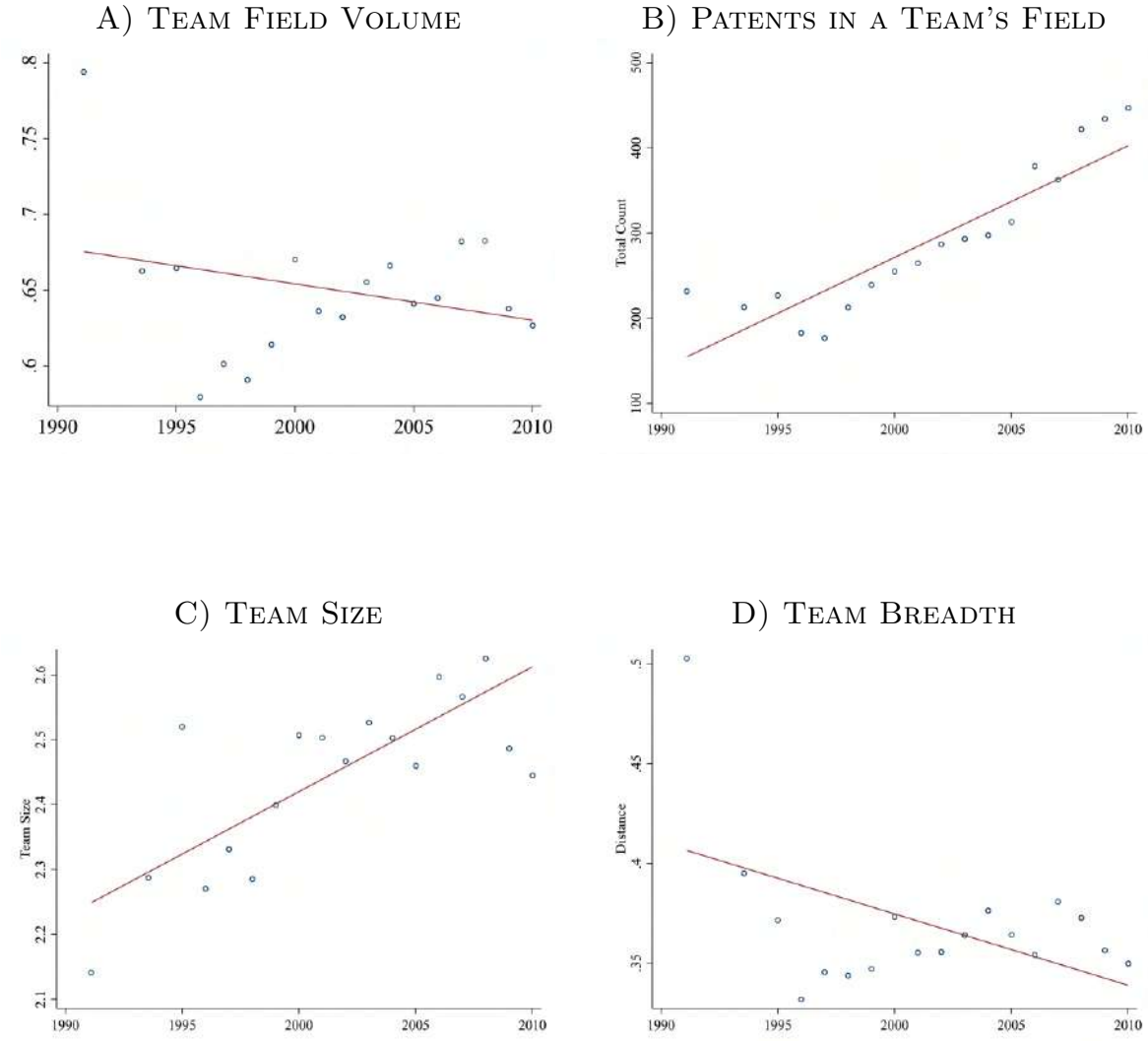
Notes: Learnt α Dirichlet prior from the `gensim` package option *auto*. The Y-axis presents the 5 words with the largest weight within the knowledge class to word distribution for that class. The height of the bar represents the weight on that class in the Bayesian prior.

Figure A5

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 Cancer Moonshot: USPTO (2024). Again the top 5 words per class shown on the Y-axis.

Figure A6
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 **??** however summing over the team field instead of a patents. Volume is defined in equation 9 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.

Table A1
PATENT REGRESSION ESTIMATES: DIRECTION

	Dependent variable: Pr(Direction)			
Prior work | Direction$_{pt}$	0.0290***	0.0281***	0.0281***	0.0438***
	(41.51)	(44.80)	(44.88)	(47.19)
Prior work | Direction$_{pt}$ Sq.				-0.0001***
				(-24.85)
N	1218385	1218385	1218366	1218366
Controls	✓	✓	✓	✓
Direction FE	✓	✓	✓	✓
Year × Direction FE		✓	✓	✓
Team size			✓	✓

Notes: Each column corresponds to a logistic regression of the probability a patent is one of three types (z), where all three types are stacked into one regression model. 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 $d(\theta_p^e, \theta_p)$.

Table A2
PATENT REGRESSION ESTIMATES: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
Prior work$_{pt}$	-0.0019***	-0.0014**	-0.0014**	0.0008
	(-3.76)	(-2.88)	(-2.89)	(1.60)
Prior work$_{pt}$ Sq.				-0.0000**
				(-2.97)
N	408772	408772	408753	408753
Controls	✓	✓	✓	✓
Year FE		✓	✓	✓
Team size			✓	✓

Notes: Each column corresponds to a logistic regression of the probability a patent is either a breakthrough. The dependent variable is the probability that patent is in the top 75% of the breakthrough score. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). All standard errors are clustered at the knowledge cluster × year level. Controls include $d(\theta_p^e, \theta_p)$.

Table A3
TEAM TREATMENT ESTIMATES: HETEROGENOUS

I: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
	1.	2.	3.	4.
$D_{\tau t}$	-0.0078*	-0.0033*	-0.0011	0.0053***
	(-2.19)	(-2.58)	(-1.06)	(6.13)
$n_{\tau t}$	-0.1835**	-0.0581***	-0.0215**	-0.0101***
	(-3.29)	(-3.32)	(-2.73)	(-4.95)
Volume_{τ}	-3.5505***	-3.0948***	-2.5336***	2.0023
	(-4.69)	(-4.85)	(-4.54)	(1.88)
N	2735	2509	1907	1659
Controls	✓	✓	✓	✓
Team FE	✓	✓	✓	✓
Period FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓

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. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Controls include $d(\theta_p^e, \theta_p)$.

B Hypothesis Development

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 $y_p \in \{b_p, z_p\}$

$$y_p(\theta_p^e, t) = \begin{cases} 1 & \text{if } f_y(\theta_p^e, t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

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. I will now link this set-up to both hypotheses outlined in section 2. These therefore target whether a patent is a breakthrough, or not. However a similar argument can be easily derived for whether a patent targets a given direction, as discussed in section 6.1.

Hypothesis 1 proposes that there exists an inverted-U shape relationship between the probability a patent is a breakthrough and the quantity of prior work on which it builds. Formally this is given by

$$\frac{\partial f_b}{\partial n_{pt}} > 0 \quad \text{and} \quad \frac{\partial^2 f_b}{\partial n_{pt}^2} < 0$$

.

This can be tested directly in the data through a logit model that captures the latent variable structure. Given the definition of a team span in equation 4, we can define the expected value for each outcome as the following.

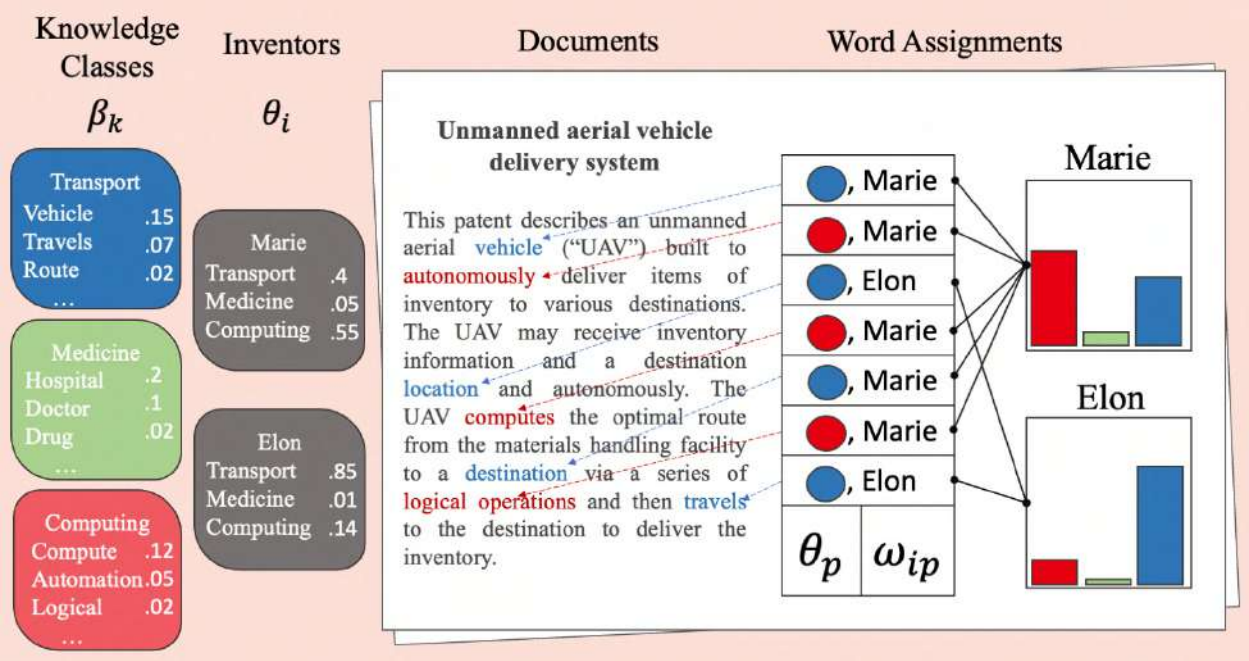
$$\mathbb{E}[y_p|\tau, t] = \frac{1}{\text{vol}(S(\tau))} \int_{S(\tau)} y_p(\theta_p^e, t)\, d\theta_p^e \tag{15}$$

In other words, what proportion of all the teams potential ideas achieve outcome y? Each potential project can be given a probability of being a breakthrough or not, through the latent variable model outlined. Therefore since all projects are drawn with a uniform probability, I can test the expected team patent outcome again using a logit model. Thanks to the uniform distribution assumption, the expected value defined in equation 15 relies on

the volume of a team span. Therefore hypothesis 2 utilises a change in the density of patents within a team's field to distinguish each case.

C Estimating the Knowledge Space: Intuition

Figure A7
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.

D 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 distance 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.

E 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, and Jordan, 2003. The Author-Topic-Model was first introduced by Rosen-Zvi et al., 2012. This is replication of the model in Mortensen, 2017 where I have simply adapted the notation from the original papers to the context of an Inventor-Knowledge Class-Model, where inventors write patent texts collaboratively.

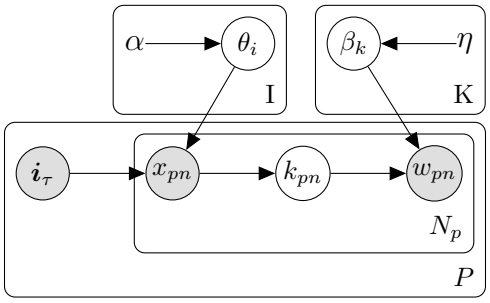
A patent p is a vector of N_p words $\boldsymbol{w}_p$ where each word w_{ip} is chosen from a vocabulary of size V, and a vector of n_τ inventors $\boldsymbol{i}_\tau$. A collection of P patents is therefore defined as $\mathcal{P} = \{(\boldsymbol{w}_1, \boldsymbol{i}_1, \cdots, (\boldsymbol{w}_P,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

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

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

This model is represented in the following plate diagram in figure (18).

Figure A8
INVENTOR-KNOWLEDGE CLASS MODEL



Notes: Plate notation for Bayesian Hierarchical model.

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

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

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

$$q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\lambda, \gamma, \phi) = q(\boldsymbol{\Theta}|\boldsymbol{\gamma})q(\boldsymbol{\beta}|\lambda)q(\boldsymbol{k}, \boldsymbol{i}|\phi) \tag{17}$$

$$\approx P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\boldsymbol{w}, \alpha, \eta, \boldsymbol{T}) \tag{18}$$

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 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 $q(\cdot)$

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

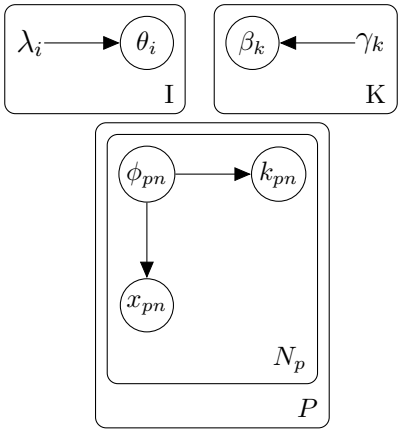
Which is the product of the probability of observing I individual knowledge class distributions, K knowledge class to word distributions and a set of inventor and knowledge class combinations for each word of every patent.

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 (19) presents the final model.

For a given patent p the matrix ϕ_{ik} gives the discrete joint probability of choosing each

[22] 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.

Figure A9
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.

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}[23]. 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[24]

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

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

[23]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.

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

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 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

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

F 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.

Table A4
Team Treatment Estimates: Kelly et al., 2021

I: Breakthrough

	Dependent variable: Pr(Breakthrough)			
	1.	2.	3.	4.
$D_{\tau t}$	0.0013***	0.0025***	0.0014***	0.0014***
	(9.18)	(6.60)	(4.38)	(4.37)
Prior work$_{\tau_1 t}$	-0.0016***	-0.0146***	-0.0045***	-0.0045***
	(-14.79)	(-17.50)	(-7.42)	(-7.43)
Volume$_{\tau}$				0.1264
				(0.44)
N	30824	10553	10553	10553
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with 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 is an indicator for whether the patent is a breakthrough using the Kelly et al., 2021 data. Controls include $d(\theta_p^e, \theta_p)$.

Table A5

TREATMENT TEAM ESTIMATES: ADDING AN INVENTOR

I: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
	1.	2.	3.	4.
$D_{\tau t}$	-0.0027**	-0.0031*	0.0002	-0.0002
	(-2.76)	(-2.30)	(0.26)	(-0.24)
Prior work$_{\tau_1 t}$	-0.0008***	-0.0682***	-0.0213***	-0.0211***
	(-10.72)	(-22.92)	(-9.78)	(-9.67)
Volume$_{\tau}$				0.4732
				(1.33)
N	33958	11069	11069	11069
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

II: DIRECTION

	Dependent variable: Pr(Direction)			
	1.	2.	3.	4.
$D_{\tau t}$ | Direction	0.0345***	0.0270***	0.0121**	0.0130*
	(8.23)	(5.62)	(2.58)	(2.51)
Prior work$_{\tau_1 t}$ | Direction	0.0122***	0.0525***	0.0254***	0.0254***
	(10.12)	(38.70)	(27.99)	(27.94)
Volume$_{\tau}$				-0.1258
				(-0.48)
N	33207	25734	25734	25734
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period × Direction FE		✓	✓	✓
Year × Direction FE			✓	✓

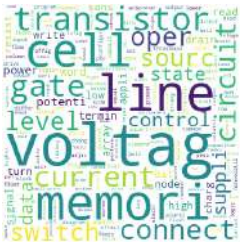
Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable for panel I) is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). 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 $d(\theta_p^e, \theta_p)$.

G Wordclouds

These are the fifty word clouds, one for each of the estimated knowledge classes, in addition to figure 14, 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 un-supervised 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.

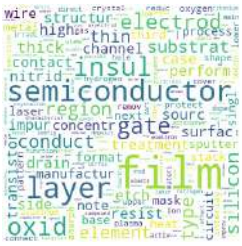
Figure A10
KNOWLEDGE CLASS TO WORD DISTRIBUTIONS: WORDCLOUDS

16)



17)

18)



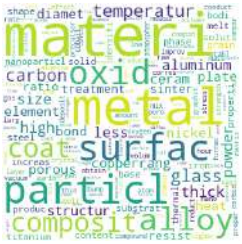
19)

20)

21)

22)

23)



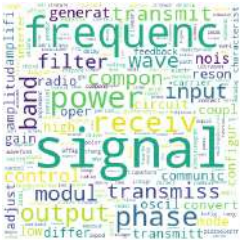
24)

25)

26)

27)

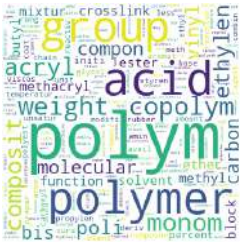
28)



29)

30)

31)

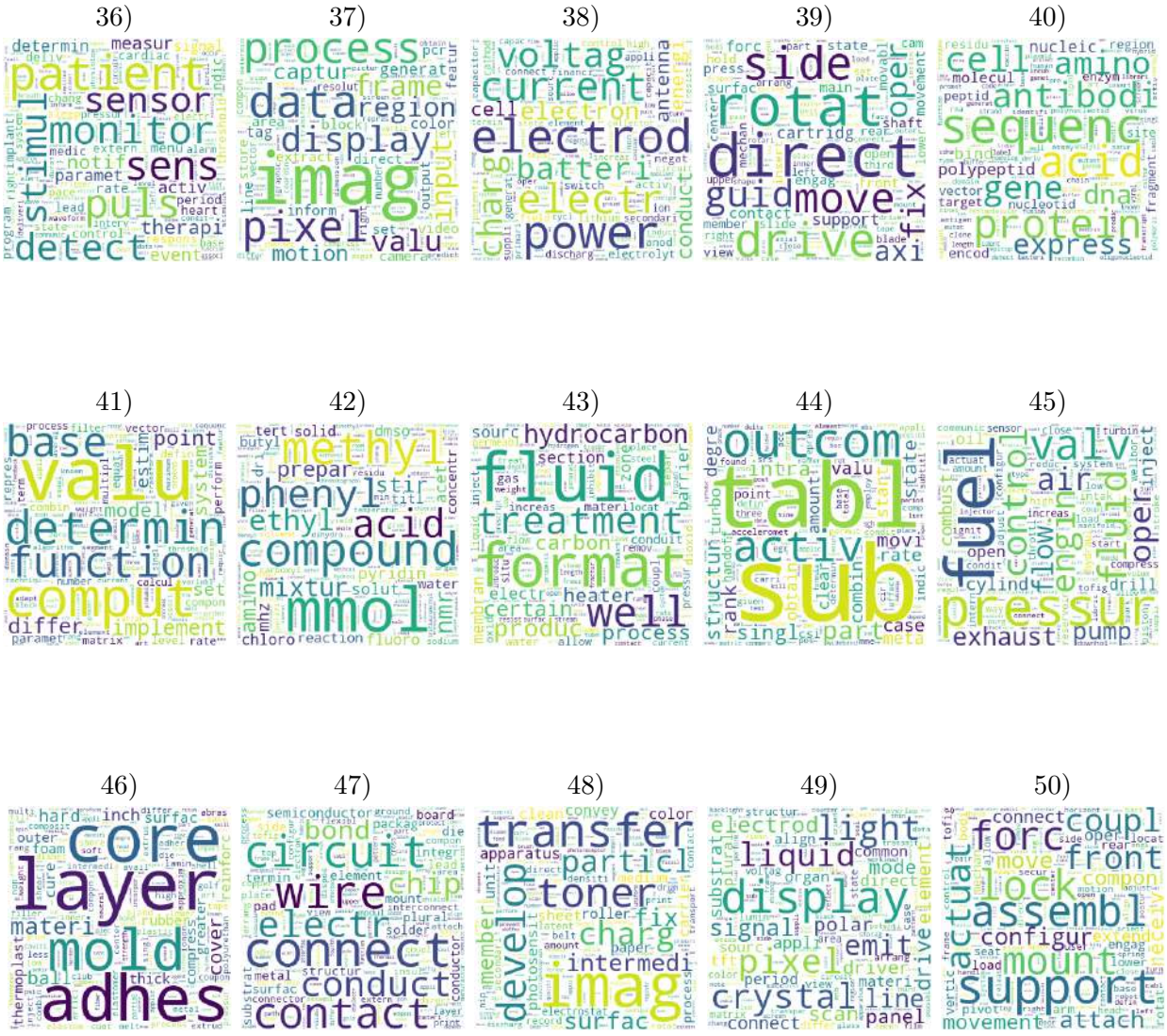


32)

33)

34)

35)



Notes: Each of the 50 knowledge class to word distributions represented as word clouds. 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.