Design Science

Inventors’ Explorations Across Technology Domains

Jeff Alstott*[1,2], Giorgio Triulzi*[1,3,4], Bowen Yan[1], and Jianxi Luo[1]

[1] *Singapore University of Technology and Design*
[2] *Massachusetts Institute of Technology, Media Lab*
[3] *Massachusetts Institute of Technology, Institute for Data, Systems, and Society*
[4] *United Nations University - MERIT*
* *These authors contributed equally to this work.*

Abstract

Technologies are created through the collective efforts of individual inventors. Understanding inventors’ behaviors may thus enable predicting invention, guiding design efforts or improving technology policy. We examined data from 2.8 million inventors’ 3.9 million patents and found most patents are created by “explorers”: inventors who move between different technology domains during their careers. We mapped the space of latent relatedness between technology domains and found explorers were 250 times more likely to enter technology domains that were highly related to the domains of their previous patents, compared to an unrelated domain. The great regularity of inventors’ behavior enabled accurate prediction of individual inventors’ future movements: a model trained on just 5 years of data predicted inventors’ explorations 30 years later with a log-loss below 0.01. Inventors entering their most related domains was associated with patenting up to 40% more in the new domain, but with reduced citations per patent. These findings may be instructive for inventors exploring design directions, and useful for organizations or governments in forecasting or directing technological change.

Introduction

Technology development and engineering design have been characterized as processes of creative transformation, recombination or synthesis of prior technologies and related knowledge into new creations (*1–5*). Highly-novel designs as well as knowledge production have been shown to arise from creative interdisciplinary combinations (*3*, *6–10*). These empirical evidences have influenced innovation policies, which are increasingly encouraging inventors to develop a broad knowledge base and to span across technology domains (*11–13*). However, little is known about how inventors explore across technology domains during their design efforts, such as whether there are typical patterns of exploration or patterns associated with higher outcomes. Understanding the general trends of inventors’ movements and performance across domains may allow for better prediction of future invention, facilitate design-by-analogy inventive efforts (*14–16*) and enable more fine-grained innovation polices.

Do inventors preferentially explore domains of technology closely related to what they already know? What are the performance implications of movements into related versus unrelated domains? We can now address these questions thanks to accurate tracking of inventions’ inventors (*17*) and quantitative measures of the relatedness between technology domains (*18–21*). The relatedness of technologies and products is important for the outputs of firms (*22–25*), cities

Corresponding author
Giorgio Triulzi
gtriulzi@mit.edu

Des. Sci.
journals.cambridge.org/dsj
DOI

the Design Society
a worldwide community

Cambridge
University Press

(*26*), regions (*27–29*) and countries (*30*): they are more likely to successfully generate an output in a new domain if it is highly related to their previous domains. All of these entities, however, are composed of individual inventors, which have greater cognitive constraints than the organizations they make up. Here we study at scale how individuals' inventive behaviors are shaped by the relatedness of technology domains.

We used data from 3,910,549 patents awarded by the United States Patent and Trademark Office between 1976 and 2010 to track 2,756,382 inventors as they patented their inventions. We found that inventors were far more likely to enter domains related to their previous work, and when they did so they patented more in the new domain. The great regularity of inventors' explorations across domains made their movements predictable, which enabled us to further create an accurate predictive model of inventors' future exploration, and allows for more informed interdisciplinary design efforts.

Aims

Here we aim to measure, at scale, inventors' behavior as they explore different technology domains. We seek to find a quantification of how much technology domains relate to each other, and then use that information to build a predictive model of inventors' movements across domains. We will also evaluate if different types of movements are associated with different performance outcomes, using several dimensions of performance.

Literature Review

Technological innovation is the process of producing effective solutions to engineering design problems. This process' characteristics have been studied by scholars of design science, creativity and innovation. These different branches of study have repeatedly concluded that high-impact inventions are generated from the creative combination of existing solutions (*2–5*, *7*, *9*, *31*, *32*). The broad agreement on this view of the engineering design and innovation process has lead to the formulation of national and supranational science and technology policies that stimulate inventors to source knowledge from different technology and scientific domains (*11–13*).

Combining prior knowledge from existing engineering solutions may not be easy, and the difficulty depends on the strength of the relatedness between different pieces of knowledge. We can characterize different technologies to be more or less similar in terms of how much their inventive process shares similar knowledge inputs and capabilities (*18*). Inventors can leverage existing knowledge and existing design solutions from close and distant technology domains. Large-scale studies using data from patents and publications have shown that high-impact ideas usually come from domain-spanning work, often with the help of teamwork by groups of inventors (*9*, *33*). Uzzi et al. (*9*) showed how atypical combinations of a few rarely co-cited scientific publications (often from different fields) and often co-cited papers (typically from the same field), increases the probability that the focal publication will be in the top 1% of most cited publications. Other studies analyzing patent data broadly confirmed these findings (*10*, *34*, *35*). This suggests that high-impact ideas are unusual in that they combine popular conventional knowledge with pieces of knowledge that are usually consider to be cognitively unrelated. However, it is important to distinguish between the potential value of combining different knowledge and its actual realization. Using patent data, Fleming (*7*) showed how experimentation with new components of technical knowledge or new combinations of different pieces of knowledge leads to less useful inventions on average, but it also implies an increase in the variability that can result in both failure and breakthrough.

Studies of engineering design provide more specific insights on how designers source knowledge to create novel artifacts. Designers' work is strongly guided by their prior knowledge and experience (*4*, *31*, *32*). On the other hand, designers also take inspiration from existing analogous solutions (*6*, *36*, *37*). However, designers' ability to mix very different strands of knowledge is constrained by their experience and their understanding of a set of scientific principles specific to their own field of work. Therefore, inventors are usually better equipped to incorporate existing knowledge from relatively similar analogous solutions in their creative process (*14*, *15*). In a recent text analysis of design concepts from a Web-based innovation platform, Chan et al. (*36*) found that conceptually closer, rather than farther, sources of inspiration lead to more useful and appropriate ideas. However, several other case studies of specific design processes have shown how moderate (*1*) or even distant knowledge inputs (*38–40*) can result in particularly novel ideas. For instance, Fu et al. (*1*) performed an experiment in which patents that they classified as near or far analogous solutions to a given design problem were provided to designers as creative stimuli. They found that designers who were exposed to existing solutions from "far" fields performed poorly, being unable to effectively integrate them into the design process. Similarly, designers who were only given "near" patents produced design solutions with limited novelty, albeit with higher quality than the former. This shows how near fields are perceived to be more relevant sources of knowledge for the design process. However, the authors also discussed how stimuli of a moderate distance may be most conducive to the successful generation of highly novel solutions. Following this idea, Fu et al. (*41*, *42*) developed a method to measure how analogous two patented design solutions are to each other to help designers identifying the most useful stimuli for their creative process. The effect on design output of distance of external stimuli to the design problem were also studied by Chan et al. (*43*) through a designed experiment similar to (*1*). They found a positive effect of far and less-common stimuli on novelty but also on the variability of a solution quality.

Taken collectively, these studies of engineering design suggest that, despite inventors' natural cognitive tendency to build on their prior experience and knowledge, designers may actually benefit from seemingly unrelated solutions. Mixing this new knowledge with what they have learned in the past in their own inventive history can lead to higher-novelty creations. This is likely to lead to failure in many instances, but a few solutions may prove to be breakthroughs (*44*). Designers can therefore benefit from tools that suggest to them where these potentially useful analogous existing solutions can be searched. One such tool could be a large-scale map of the space of technology domains and how closely they relate to each other. Such a map would help designers better understand how different knowledge can be mixed, foster diversification of inventive output and facilitate design-by-analogy.

The importance of mapping the technology space has lead to multiple attempts to use patent data to describe the space. Multiple measures have been developed to quantify how technology domains relate to each other. However, different measures produced different maps (*19*, *20*, *45*). We have recently shown that all the most popular measures of relatedness are affected by several confounding factors (such as different domains having different ages or very different numbers of patents). When these factors are controlled for the different maps all collapse into much closer agreement, and these maps are on the whole very stable. (*18*). Here we will use these methods to quantify how hundreds of domains relate to each other, and use that information to predict inventors' explorations across the space of technology domains.

A similar question has been addressed in the context of firms' exploration and diversification into new product areas. Several studies showed that firms tend to preferentially diversify into related domains, though they can enter less related domains as well (see, for instance, (*8*, *22–25*)). This is likely because firms' explorations are much less constrained than individuals, most obviously through greater financial resources. However, firms can also have greater cognitive capacity, firstly by having multiple minds in the form of different employees, but also by the ability to simply acquire existing firms in operating in different domains. On a methodological point, previous studies on firm behavior have also used imperfect metrics of relatedness that may affect the correct inference of exploration behaviors, as shown by (*23*). Hence, the present work addresses an important methodological and research gap on inventor's search strategies and contributes to the design science community by examining the behavior of individual inventors, instead of management and organizations.

Methods

Data

Data on all patents granted between 1976 and 2010 by the United States Patent and Trademark Office (USPTO) were acquired from the USPTO's public data sets hosted by Google at https://www.google.com/googlebooks/uspto.html. Each patent contained three pieces of metadata used in the present analysis:

1. The domain of the invention (its classification in the International Patent Classification system, at the 4-digit level, which has 629 classes.)

2. Citations to other patents (if any)

3. Who invented the invention (the name(s) of the patent's author(s), disambiguated with data from (*17*))

More information about and interpretations of these metadata are included in *Appendix*.

Measuring Technology Relatedness by Comparison to Random Expectation

We sought to predict what domain an inventor would explore next by quantifying how related other domains were to the domains of the inventor's previous patents. We called two domains "related" if they had an unusual amount of interactions in the patent record. Domains' patents can interact in many ways, such as by citing each other; such citations are a signal of technical proximity or knowledge coupling between domains (*21*, *46*, *47*). Thus we intend to measure the relatedness between two domains by how much patents from the two domains interacted with each other, such as by citing each other. However, citations (and many other ways that patents or domains can interact) are affected by more than just the relatedness between domains. As a simple example, if the domains both had many patents we would expect a large number of citations simply by chance. As a more complicated example, the age distributions of patents in each domain also affects citation rates between them, due to the peculiar shape of the age distribution of citations and the increasing average number of citations made by a newly granted patent, as shown in (*18*, *46*, *48*, *49*). These are examples of what we have called "impinging factors", which affect the measured interactions between technology domains but are not representative of the relatedness of technology domains.

We previously showed how to normalize several different measures of relatedness by controlling for many impinging factors at once (*18*). This was done by comparing the empirical data of interactions ($I_{empirical}$) to what the data would be expected to look like by chance, holding several factors constant. To calculate how much two domains are expected to interact ($I_{expected}$), given observed properties such as their number of patents, we used link swapping to create 1,000 randomized versions of the historical record of the nearly 4 million patents. These randomized versions of history preserved the following features:

1. the number of patents in each domain
2. each patent's number of citations, both made and received
3. each patent's portion of citations to patents in other domains ("cross-domain" citations), both made and received
4. each patent's exact citation age structure (e.g. a citation to a patent of granted in year 1980, and a citation received from a patent granted in 2002)
5. the age structure of each patent's cross-domain citations (e.g. the citation to a patent in 1980 was cross-domain, and the citation received from a patent granted in 2002 was same-domain)

For each of these 1,000 versions of history we calculated the number of citations patents in each domain make to patents in each other domain. We could then measure if two domains interact more than the expectation: $I_{empirical} > I_{expected}$. Patents have numerous kinds of interactions to measure, but their deviations from expectation correlate: different measures give similar stories of how much two domains are related (*18*). Here we measured interactions ($I_{empirical}$ and $I_{expected}$) simply by how much two domains' patents cite each other, using data from 35,129,936 citations (results using other measures of interaction are qualitatively similar and are shown in *Appendix*).

If domains cite each other more or less than expected, this could be due to noise: the expectation $I_{expected}$ has variance, as could the influence of latent relatedness. We increased our confidence that two domains were related through repeated samples: using patents awarded each year from 1976 to 2010 we counted how many years $I_{empirical}$ was greater than $I_{expected}$. We expressed this

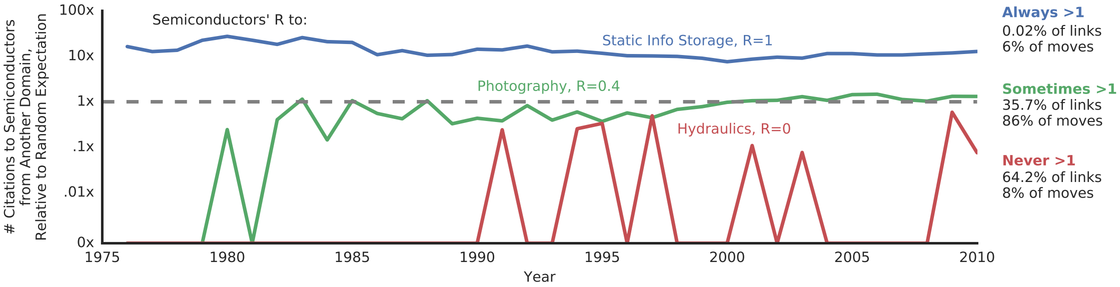


Figure 1. Three examples of how relatedness, *R*, was calculated. Each year citations to patents in the Semiconductors domain from patents in other domains were counted and compared to the quantity expected by random chance, given all domains' number of citations and other factors. *R* was the portion of years that the number of citations was above expectation.

count as a percentage, *R*, which was our measure of relatedness (Fig. 1). Pairs of domains with high *R* persistently interacted more than expectation, and were interpreted as more likely to be related. The majority (64%) of domain pairs always had fewer citations than expectation ($R = 0$), such as "semiconductors" to "hydraulics" (Fig. 1, red line). Just 0.02% of domain pairs always had more citations than expectation, such as "semiconductors" to "static info storage" (Fig. 1, blue line).

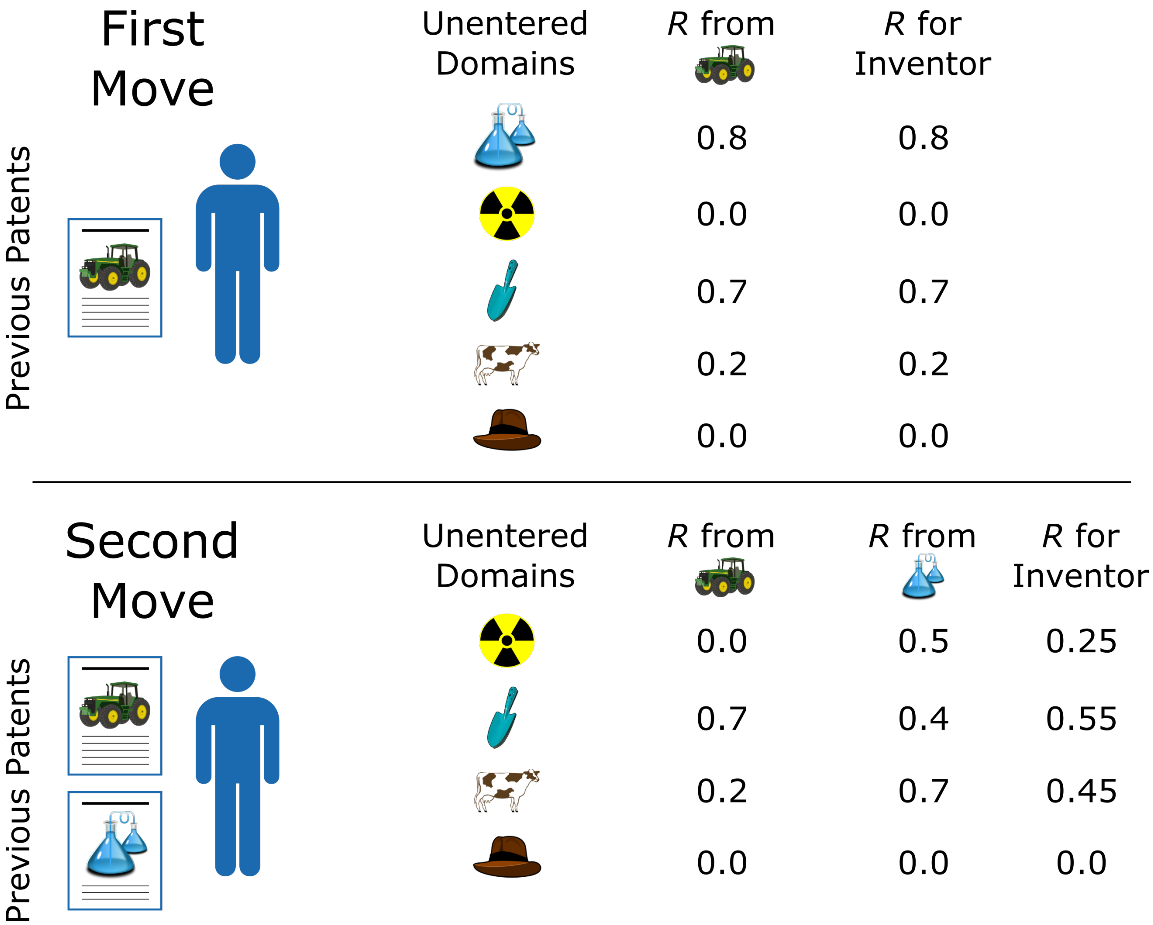
Measuring the Relatedness of an Inventor's Previous Domains to Unentered Domains

With a measure of the relatedness between domains in hand, we then quantified how much an inventor's existing knowledge was related to any specific domain. This was done through simply identifying every domain in which the inventor had previously patented, and then taking the mean R between those domains and the specific domain in question (Fig. 2). This quantity was calculated for every domain the inventor had not patented in. This quantity was calculated initially only accounting for the inventor's first domain, and then recalculated every time the inventor patented in a new domain. This mean R was the measure used for models predicting inventors' movements and performance (see below). Note that for these models the R used was the R as measured in the year immediately before the inventor filed their patent in a new domain, and so it did not include information about the inventor's new patent, or indeed any future patents.

Predictive Model

Each time an inventor patented in a new domain we sought to predict what the domain would be. We created a predictive model that relied on the following data:

- the mean R of the entered domain to the inventor's previous domains
- the popularity of the entered domain (the number of patents granted in the domain in the year before the inventor patented in it)
- whether the inventor's previous patents had cited patents in the entered domain



Unentered Domains	R from 🚜	R for Inventor
⚗	0.8	0.8
☢	0.0	0.0
🪏	0.7	0.7
🐄	0.2	0.2
🎩	0.0	0.0

Unentered Domains	R from 🚜	R from ⚗	R for Inventor
☢	0.0	0.5	0.25
🪏	0.7	0.4	0.55
🐄	0.2	0.7	0.45
🎩	0.0	0.0	0.0

Figure 2. An inventor's R to an unentered domain was the average of the R to the domain from each of the domains in which they had previously patented. Top) A diagram of an inventor before their first move. Since the inventor had patented in only one domain, the inventor's R to each unentered domain was the same as the R from that one domain. Bottom) A diagram of the same inventor before their second move. At this point the inventor had patented in two domains, so the inventor's R to each unentered domain was the mean of the Rs from the two domains. Actual technology domains were very fine-grained (examples shown in Fig. 10).

- whether the inventor's previous co-authors had patented in the entered domain

Aside from R, the other factors are clearly potentially relevant to an inventor's behavior. The popularity of the entered domain is meant to control for both the probability that an inventor's patent is classified (or misclassified) into a large technology class just by random chance, as well as account for the possibility that inventors chase hot topics, regardless of its relatedness to their previous work. Information on the inventor's co-authorship network is included to account for the possibility that the inventor had prior personal connection with authors in a newly entered domain, which may facilitate exploration. Empirical evidence that research is increasingly done in teams (*33*), has led to speculations of a growing need to rely on inter-domains knowledge across team members to achieve high impact (see for instance (*50*) and (*51*)). This may motivate inventors to enter domains in which they know previous co-authors. The inventor's previous citation(s) to a domain seeks to measure the existence of a personal knowledge bridge between the inventor's prior work and a domain, which could influence the inventor's exploration decisions more than the overall relatedness between domains. Such a personal bridge may exist due to several unobserved factors, such as the inventor's educational background or personal connections with colleagues that have not co-authored the inventor's previous work, as suggested by (*52*)

The predictive model was created by identifying all the domains the inventor had not yet entered, then calculating $p(\text{entry})$ for each domain using a naive classifier:

$$\begin{aligned} p(\text{entry}) \quad &\sim \quad p(\text{entry}|R) * p(\text{entry}|\text{popularity}) \\ &\qquad * p(\text{entry}|\text{co-authors}) * p(\text{entry}|\text{citations}) \end{aligned} \tag{1}$$

$p(\text{entry})$ for each domain was thus a function of its mean R to each of the inventor's previous domains, the popularity of the domain, whether the inventors had co-authors previously active in the domain, and whether the inventor's previous patents had citations to patents in the domain. For each of these variables x we estimated $p(\text{entry}|x)$ by simply creating a histogram from historical data (this created a discrete naive classifier, discussed further in *Appendix*). For co-authors and citations this histogram had only two bins: ($p(\text{entry}|\text{had co-authors active in domain})$ vs. $p(\text{entry}|\text{no co-authors active in domain})$) and ($p(\text{entry}|\text{had citations to patents in domain})$ vs. $p(\text{entry}|\text{no citations to patents in doma}$ For popularity, we created a histogram with 500 bins (from popularity percentile rank 0 to rank 100, in intervals of 0.2%). Similarly, for R we created a histogram of with 26 bins ($R = 0$, plus 25 bins evenly spaced up to $R = 1$). $p(\text{entry}|x)$ was then taken to be the $p(\text{entry})$ for the bin that x was in. We calculated $p(\text{entry}|x)$ for each year individually, using data from 1976 up through that year. Each inventor exploration was predicted using data only up through the year before the year the inventor applied for a patent in a new domain.

Performance Models

Given that the next domain to enter could be predicted, one would be further interested in predicting the inventor's performance in the new domain. We modeled inventors' performance as patent and future citation counts when they patented in a new domain as a function of the properties of the domain and each individual inventor's history. These properties included:

- the mean R of the entered domain to the inventor's previous domains
- the popularity of the entered domain (the number of patents granted in the domain in the year before the inventor patented in it)
- the number of times the inventor's previous patents had cited patents in the entered domain (if any)
- the number of times the inventor's previous co-authors had patented in the entered domain (if any)
- the inventor's previous rate of producing patents (which has been shown to correlate with inventor's future patenting rates (*53*))

These properties were combined into a vector, $\vec{x}$, and the outcomes were modeled as a function of this vector, $f(\vec{x})$.

Patent and citation counts were modeled as generated from a negative binomial distribution, a classic count model with a variance that can be much larger than the mean. The negative binomial had the form:

$$p(\text{count}|\mu,\phi) = \binom{\text{count}+\phi-1}{\text{count}}\left(\frac{\mu}{\mu+\phi}\right)^{\text{count}}\left(\frac{\phi}{\mu+\phi}\right)^{\phi} \tag{2}$$

where μ is the mean and ϕ is the overdispersion parameter (which determines the variance). We set μ as a function of $\vec{x}$ and a baseline outcome: the average number of patents or citations received by all explorers who entered the same domain in the same year. This function had the form $\mu = e^{\beta\vec{x}} * \text{baseline}$. Thus, for each property x_i in $\vec{x}$, β_i reflected how much x_i was associated with increasing or decreasing performance relative to the baseline of other explorers.

We used Bayesian inference to find the most credible values of the parameters β and ϕ, given the data and Gaussian priors for both (β priors: normal distributions of mean 0 and standard deviation 2. ϕ prior: a gamma distribution with parameters μ_ϕ^2/var_ϕ and μ_ϕ/var_ϕ, where μ_ϕ and var_ϕ also had normal priors of mean 0 and standard deviation 2). Fig. A3 shows the resulting posterior distributions for ϕ and for each β_i for each property x_i. These posteriors were calculated using Hamiltonian Monte Carlo sampling, as implemented in the software package *Stan* (*54*). We sampled the posteriors with 50 chains of 300 iterations of warm-up and 300 iterations of sampling, thinned down to 500 uncorrelated samples in the posterior.

To assess if these count models were well-specified to the data, we used posterior prediction for each entry to calculate the models' 95% credible interval for the explorer's number of future patents or citations in the domain. The observed patent and citation counts were within the models' 95% credible interval approximately 95% of the time (patents: 97.7%, citations: 96.6%).

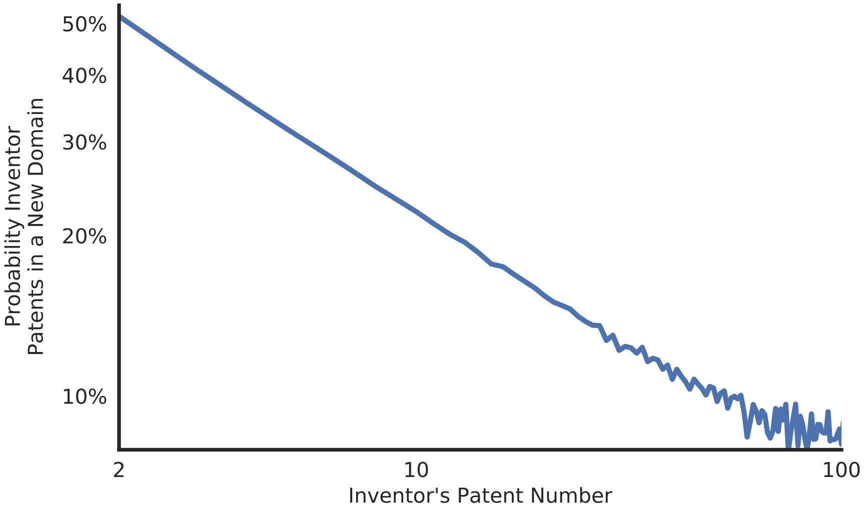


Figure 3. Inventors regularly explore across technology domains. The probability that an inventor's next patent was in a previously unentered technology domain, given the number of patents the inventor already had (double logarithmic axes).

When analyzing and modeling performance we only included the entries up to 2005 (as opposed to 2010), to ensure each entry had at least 5 more years to observe the inventor's subsequent performance.

Data Availability

All data and code for analyses, including generating figures, is contained in the Supporting Information. All code is also hosted at https://github.com/jeffalstott/inventorexploration.

Results

Most (60%) inventors patented only once, but 84% of patents were made by repeat inventors with more than one patent. Inventors' patents were classified into 629 domains of technology by domain-expert patent examiners (see *Appendix*). We could thus track as repeat inventors patented in one technology domain, such as "semiconductors", and then later patented in another domain, such as "photography." We call this "exploring" or "entering" a new domain and inventors who do this "explorers." 71% of repeat inventors were explorers, and they were granted 77% of patents. Explorers made a total 1,763,920 entries, and 56% of patents were granted to an explorer who had entered the patent's domain from elsewhere. An inventor's probability of exploration was related to how many prior patents the inventor already had, with a form well-described by a power law (Fig. 3). Heavy-tailed drop-offs in humans' exploratory behavior occur in other domains (*55*, *56*) and indicate that inventors' apparent exploratory movements are not just due to erratic classification of patents (discussed in *Appendix*).

When an inventor entered a new domain, we measured the mean R from the

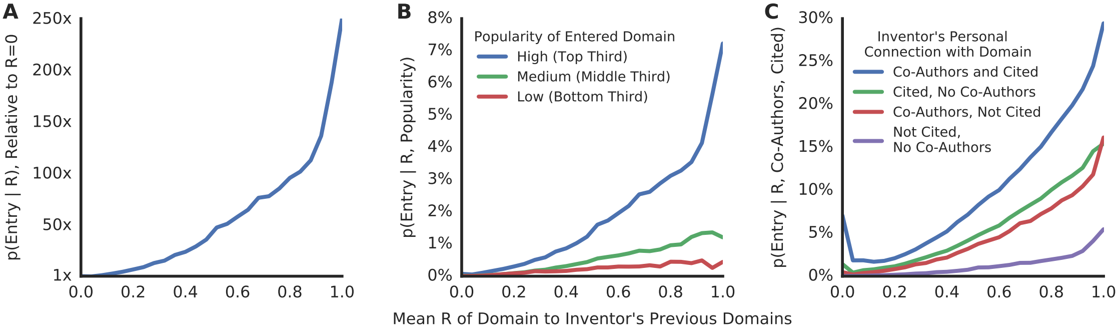


Figure 4. Inventors were far more likely to explore a domain if it was related to their previous work. A) The probability density function of how likely an inventor was to move to a domain, given its mean *R* to the inventor's previous domains. B) As A, conditional on the popularity of the domain (the number of patents in the domain in the previous year). C) As A, conditional on whether the inventor had previous co-authors who had patented in the domain before working with the inventor, and whether any of the inventor's previous patents had citations to the domain.

inventor's previous domains to each of the domains they had not yet entered (the values of *R* were calculated using only data from patents awarded up to the year *before* the inventor applied for the patent, and thus did not include data from the inventor's new invention). We found that inventors were far more likely to enter domains with high *R*; the probability of entering a domain with $R = 1$ was 250 times higher than entering a domain with $R = 0$ (Fig. 4A). Accordingly, the bulk of inventors' moves (92%) were between pairs of domains linked by $R > 0$, even though just 36% of pairs were so linked (Fig. A1; a portion of these linked pairs are visualized in Fig. 10).

The tendency to enter a domain with high R was not the only factor that predicted inventors' movements. If a domain was particularly popular, inventors were more likely to enter into it (Fig. 4B). Additionally, inventors were more likely to enter a domain if they had a more personal connection to it, such as by having had co-authors with prior patents in the domain, having had patents that cited patents in the unentered domain, or particularly having had both (Fig 4C). However, while both popularity and personal connections increased the probability an inventor would enter a domain, a higher R raised that probability further (Fig. 4B-C).

The regularity of inventors' movements allowed for predicting individuals' future explorations across domains. We applied our predictive model (a discrete naive classifier), described above, to calculate the probability of an inventor entering a domain, as a function of *R*, popularity, co-authors and citations. Each time an inventor moved we created a list of their unentered domains, ranked from most probable to least probable to be entered. Perfect prediction would always put the domain actually entered at the top of the list, and random prediction would put it in the middle of the list on average. The predictive model consistently put the domain actually entered within the top 2% of the list half the time, and within the top 7% of the list on average (Fig. 5). Other measures of predictive power include a c-statistic over 0.9 and log-loss below 0.01 (Fig. A4). Even when the model was trained only using data up to 1980, it accurately predicted inventors' movements in 2010, 30 years into the future. The persistent prediction indicates that the relationships captured in this model were stable for over 30 years. Thus,

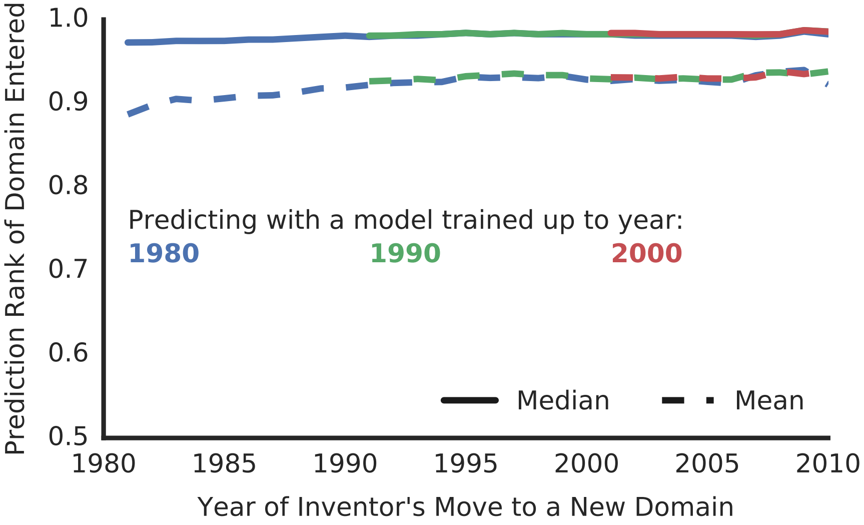


Figure 5. **Inventors' explorations were predictable.** When each inventor moved, a naive predictive model ranked the inventor's unentered domains in order of their probability to be entered. The model had been trained with patent data from 1976 up to 1980 (blue), 1990 (green), or 2000 (red). Regardless of the model used, the domain actually entered was typically high on the prediction list. Perfect prediction: 1.0, Random prediction: 0.5. Other measures of predictive power: c-statistic > 0.9, log-loss < 0.01 (Fig. A4)

the model will likely retain predictive power going forward.

Knowledge production is increasingly done by teams, and teamwork is important for combining knowledge from different fields (*9*, *33*). About half of explorers had a "guide": their first patent in the entered domain had a co-author that had patented in that domain before. Explorers were more likely to have a guide if they entered a domain with high R rather than low R (Fig. 6; 70% for $R = 1$ vs. 47% for $R = 0$). Regardless of the explorer's R to the domain entered, they were typically accompanied by 1 co-author who was also an explorer (Fig. 7). For the many explorers without guides, if the explorer had a higher R to the entered domain they were more likely to have a co-author that, while not a guide, also had a higher R to the new domain (Figs. 8, A2). Thus, those explorers entering domains with higher R were more likely to be part of teams with experience in or connection to the domain.

Where an inventor explored had ramifications for their future performance. Inventors who entered domains with high R to their previous domains went on to have more patents in the new domain (Fig. 9, blue). These patents had more total citations (Fig. 9, green), but because the number of future citations grew with R slower than the number of future patents, the average citations per patent was lower for explorers who entered a domain with higher R (Fig. 9, red). These results suggest that inventors face a performance trade-off. They are more likely to be highly productive when they explore a domain related to their prior knowledge. But the average quality of their inventions in a newly explored domain is likely to be higher if they domain is less related to their previous creations.

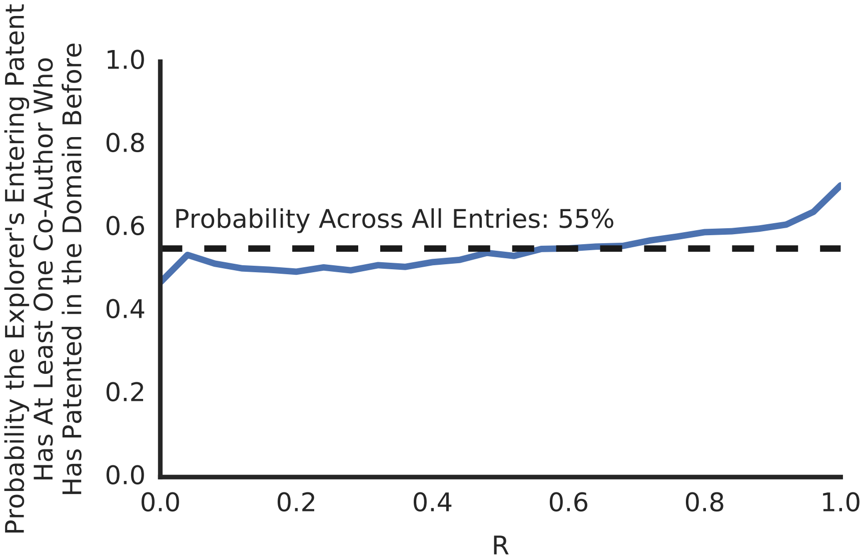


Figure 6. **Explorers were more likely to have a guide if they entered a domain with high *R* to their previous patents**.

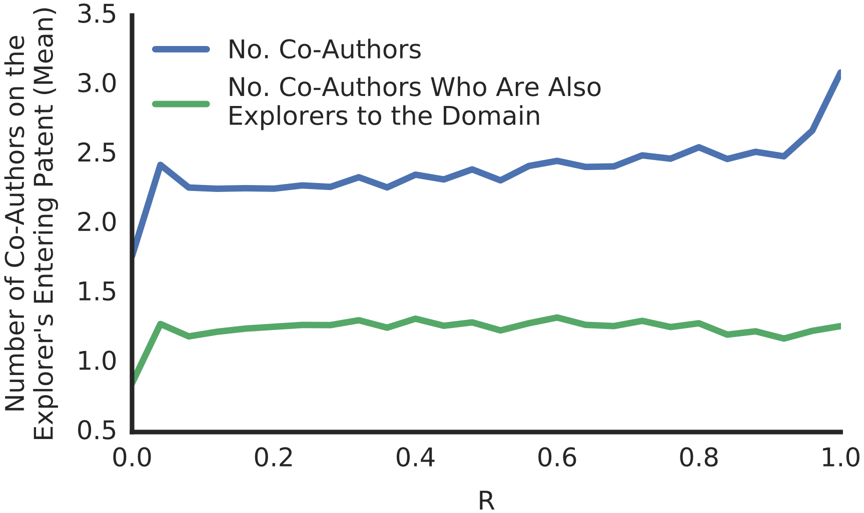


Figure 7. **The number of co-authors an explorer had on their first patent in a new domain increased with *R*, while the number of co-authors who were also explorers to the domain was largely unassociated with *R*.**

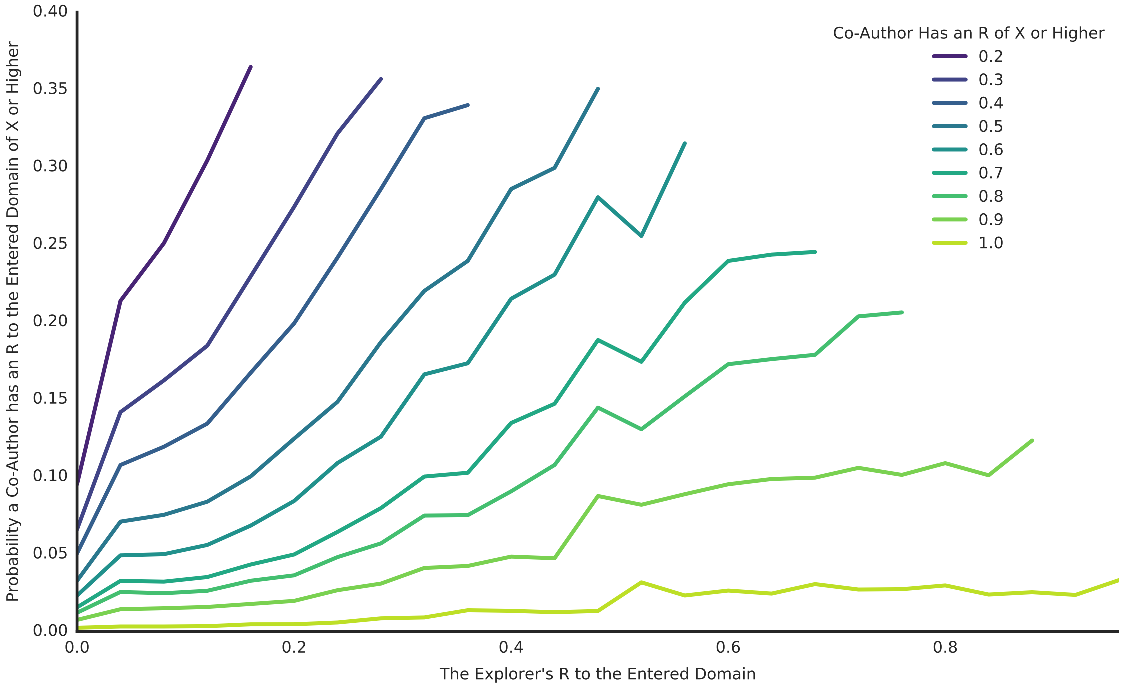


Figure 8. Explorers' first patent in an entered domain was more likely to have co-authors with higher *R* to the domain if the explorer had a higher *R* themself. When an explorer entered a new domain, their first patent in the domain could have co-authors, and those co-authors could also be explorers to the domain. Of those fellow explorers, we can ask the probability that at least one of them will have a high *R* to the entered domain, with varying thresholds for what is a "high" *R* (different colored lines). The probability of having a high *R* co-author was a function of the explorer's own *R* (x-axis). Note that the plotted lines terminate once the explorer's own *R* is above the threshold; thus, this is the probability of an explorer having a co-author with a higher *R* than their own. The data presented is for the 45% of explorers that did not have a guide on their entering patent (a co-author who had previously patented in the entered domain).

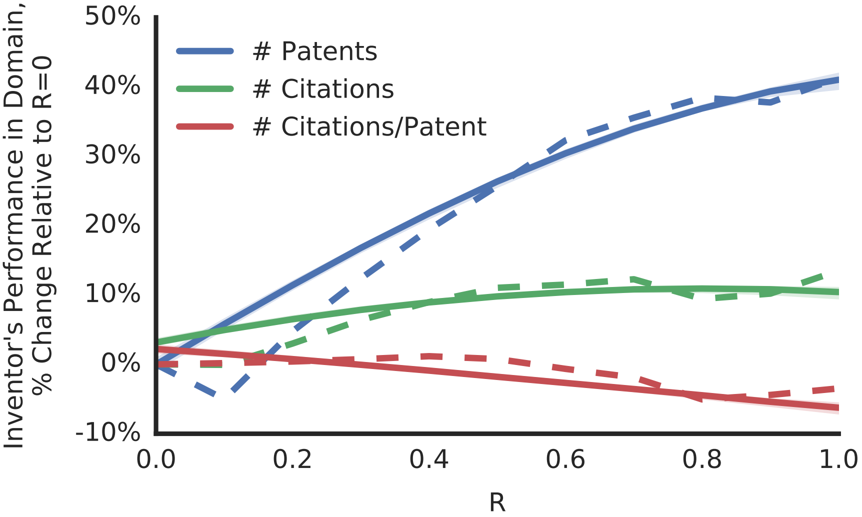


Figure 9. **Inventors had higher total performance when they entered domains more related to their previous work, but lower average citations.** Blue) An explorer's number of additional patents in the entered domain after the entering patent. Green) The number of citations those patents received Red) The average number of citations per patent. All values are expressed relative to entering a domain with $R = 0$. Dashed lines: empirical averages, binned by R in 11 bins from 0 to 1. Solid colors: association of R with performance, as inferred from models that accounted for other factors (Line: median expectation of performance, assigning all empirical entries a given value of R and holding all other parameters constant at their empirical values. Shading: 95% credible interval). Parameter values for these models are shown in Fig. A3

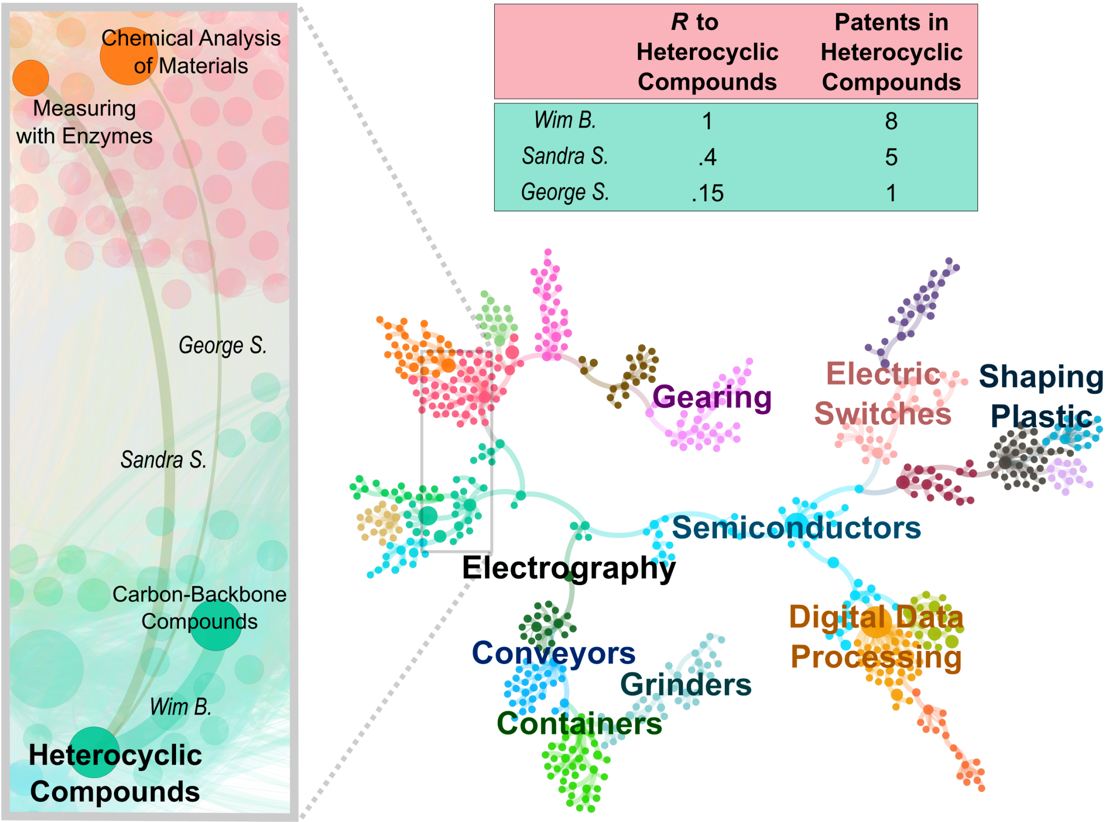
We tested if these relationships between R and performance held while accounting for other factors by applying Bayesian inference to fit the model of explorers' future performance that we introduced in *Methods*. In addition to R, this model included: the entered domain's popularity; personal connections like previous co-authors' patents in the domain and the explorer's previous citations to the domain; whether the entering patent had a guide; the entering patent's number of co-authors; the explorer's previous productivity (number of patents per year since their first patent). After accounting for these other factors, the relationships between R and performance persisted (Fig. 9, solid lines).

Discussion

We quantified the relatedness of technology domains and demonstrated its relevance to inventors' movements and performance, at scale. Using relatively simple models we were able to effectively predict inventors' explorations and describe their multi-dimensional performance outcomes. Inventors' movements across domains of technology are predictable because the moves are shaped by where the inventors have been before: domains that are related to an inventor's previous work are explored more frequently. These movements to highly related domains also yielded more patents (as was the case of inventor Wim B., Fig. 10). The entering and repeated patenting in related domains may be because related domains are comparatively easy to enter. The ease of entry likely stems from the inventor possessing relevant knowledge and skills, but it may also be due to the inventor having access to physical equipment and other external resources that affords the exploration. Exploring a less related domain is likely harder, which is why fewer inventors successfully do so, and even those who do enter don't typically patent as much (as with Sandra S. and George S., Fig. 10).

Patent counts are one measure of performance, but not all patents are equal in value: patents' citations are an indicator of how valuable the patented technology was for the economy or society, and that value likely increases superlinearly with the number of citations a patent receives (*57–61*). Inventors who explored domains with high R received more total citations, but the citations per patent were lower; depending on the exact value of citations per patent, it is thus possible that the rare inventors who successfully entered domains with lower R created inventions with more value. Creating more value from spanning disparate domains has been suggested by prior studies (*8*, *62*), but the observed increase in citations per patent after moving to lower R has a selection bias: those rare inventors who cross the chasm to unrelated domains are likely particularly talented or well-supported, and we do not observe the less-resourced inventors who tried to enter a less-related domain and failed to patent. Thus, the higher citation per patent after exploring a lower R domain could be entirely artifactual. Future studies with different data will be necessary to address this selection bias and assess the expected value of attempting to enter a less related domain.

A patent's broader impact could potentially be seen by more than just its citations received. In those rare cases where an inventor who has previously patented in a domain X creates an invention in a new domain Y that had low R with X, we hypothesize that it could create a lasting connection between those domains. This may happen when the invention created a new generic technique that reorganized the working system of an artifact, as suggested by a recent model



	R to Heterocyclic Compounds	Patents in Heterocyclic Compounds
Wim B.	1	8
Sandra S.	.4	5
George S.	.15	1

Figure 10. Three instances of inventors exploring a new technology domain. Left) In 1996 three inventors, Wim B., Sandra S., and George S., all entered the domain of "heterocyclic compounds" (chemicals with a ring of carbon and non-carbon atoms). They had all patented in only one domain previously, had no co-authors who had previously patented in the new domain, and their patents had not cited the new domain. Their performance in the new domain was related to the *R* of the new domain with their previous experience (table). Right) Maximum spanning tree of the full set of all 629 technology domains and their *R* to each other. To aid visualization, a community structure is highlighted, and some of the larger domains are labeled. Link width: *R* between two domains. The node sizes and link widths are visualized using all patent data from 1976 to 2010, but the rank order of both moved little during the years visualized.

by (*63*). A truly high-impact patent would then lead to the whole population of inventors seeing more connections between X and Y, and that this would lead to the R between the two domains to increase. Measuring such step-changes in R and their potential antecedents in individual patents is an opportunity for further research.

Predictive power's implications to practitioners and policymakers

The predictive model of inventors' explorations was so accurate over several decades because inventors' movements are very regular. A large component of this regularity is the network of relatedness between technologies, a portion of which is mapped in Fig. 10 and the entirety of which is included in *Supporting Information*. This network map and the prediction that it affords may be useful to inventors, company managers and policymakers.

The technology relatedness map and performance prediction models may aid engineering designers in their assessment of possible exploration opportunities. Designers can use the map to identify domains that they can most feasibly explore for new opportunities, given their own personal history of which domains they have prior experience in. The map can also be used to search for interesting domains that are far from a designer's previous domains. In this case, the performance models can help guide the search by estimating performance implications and trade-offs associated with exploring distant domains. Taken together, the technology relatedness map and the prediction models form a useful tool that can guide designers in their search for inventing opportunities, by helping them make more informed decisions.

Many inventors work for research labs or research divisions of companies, and in many instances an inventor's exploration of a new domain may be the result of a strategic managerial decision by the employer. However, inventors cannot be readily repurposed to new research projects in any arbitrary domain. Instead, their knowledge and experience is most readily transferable to the few related domains that are related to their previous knowledge, and those domains can be identified with the technology relatedness network map. Related domains are where inventors are more likely to be able to invent and to invent successfully, at least as measured by higher average levels of patenting. How to best pursue outlier performance, such as radical innovation, may yet require other strategies (*9*, *10*). The relatedness network map can instead be used by companies, research laboratories and research intensive government agencies to identify their workforce's expertise and where that expertise may be most readily reallocated for new projects.

Likewise, the network map of technology relatedness may also be useful for policymakers trying to foster interdisciplinary collaboration, as it can tell them which domains are likely to work well together. This may help the practical implementation of policies advocating for interdisciplinary science and technology (*11–13*). Information on which domains are related to which others may also be useful for policymakers trying to grow a target domain. The target domain could be a technology that a lagging country is trying to catch up to the level of another country; the target domain could also be a technology that is nascent around the world and the objective is to grow the domain for global benefit. In either case, it may be possible to deliberately coax inventors from related domains into the target domain to increase the manpower and invention in the technology. Whether inventors typically respond successfully to such coaxing by policymakers, or reallocation by company managers, is a question deserving further research.

The better understanding of individual inventors' behavior achieved here may also enable expanded research on principles of technology development. It would be possible to model inventor population dynamics within and across domains, or to measure how technological shocks propagate between related domains across the technology space. Understanding how individuals' choices lead to aggregate outcomes may thus be a basis for predicting broad technological change.

Acknowledgements

JA, GT and JL designed the study. BY collected the data and visualized the networks. JA and GT developed and performed the analysis. JA, GT and JL wrote the paper. We thank César Hidalgo, François Lafond, Aditya Mathur, Chris L. Magee, James Savage, Kristin Wood and Hyejin Youn for helpful discussions.

Financial Support

This work was supported by the SUTD-MIT International Design Centre (IDG31300112), the Singapore Ministry of Education Tier 2 Academic Research Grants (T2MOE1403) and the SUTD-MIT Postdoctoral Programme.

References

1. K. Fu, J. Chan, J. Cagan, K. Kotovsky, C. Schunn, K. Wood, The Meaning of "Near" and "Far": The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output. *Journal of Mechanical Design* **135**, 021007–021007 (Jan. 2013).
2. R. W. Weisberg, *Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts* (John Wiley & Sons, 2006).
3. W. B. Arthur, *The Nature of Technology: What It Is and How It Evolves* (Simon and Schuster, 2009), 256 pp.
4. A. Hatchuel, B. Weil, CK Design Theory: An Advanced Formulation. *Research in engineering design* **19**, 181 (2009).
5. V. Tang, J. Luo, presented at the International Connference on Engineering Design, ICED13, pp. 1–10.
6. Y. Reich, O. Shai, The Interdisciplinary Engineering Knowledge Genome. *Research in Engineering Design* **23**, 251–264 (2012).
7. L. Fleming, Recombinant Uncertainty in Technological Search. *Management Science* **47**, 117–132 (Jan. 2001).
8. B. Nooteboom, W. Van Haverbeke, G. Duysters, V. Gilsing, A. van den Oord, Optimal Cognitive Distance and Absorptive Capacity. *Research Policy* **36**, 1016–1034 (Sept. 2007).
9. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical Combinations and Scientific Impact. *Science* **342**, 468–472 (Oct. 25, 2013).
10. D. Kim, D. B. Cerigo, H. Jeong, H. Youn, Technological Novelty Profile and Invention's Future Impact. *EPJ Data Science* **5** (Dec. 2016).

11. G. W. Clough, *The Engineer of 2020: Visions of Engineering in the New Century* (National Academies Press, Washington, D.C., May 14, 2004).

12. S. Olson, M. Dahlberg, *Trends in the Innovation Ecosystem: Can Past Successes Help Inform Future Strategies? Summary of Two Workshops* (National Academies Press, Washington, D.C., Sept. 10, 2013).

13. "Interdisciplinarity in Research" (European Research Advisory Board (EURAB), Apr. 2004).

14. J. S. Linsey, *Design-by-Analogy and Representation in Innovative Engineering Concept Generation* (ProQuest, 2007).

15. J. S. Linsey, A. B. Markman, K. L. Wood, Design by Analogy: A Study of the WordTree Method for Problem Re-Representation. *Journal of Mechanical Design* **134**, 041009 (2012).

16. D. A. McAdams, K. L. Wood, A Quantitative Similarity Metric for Design-by-Analogy. *Transactions-American Society of Mechanical Engineers Journal of Mechanical Design* **124**, 173–182 (2002).

17. G.-C. Li, R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, L. Fleming, Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975-2010). *Research Policy* **43**, 941–955 (July 2014).

18. J. Alstott, G. Triulzi, B. Yan, J. Luo, Mapping Technology Space by Normalizing Patent Networks. *Scientometrics* **110** (2017).

19. L. Kay, N. Newman, J. Youtie, A. L. Porter, Patent Overlay Mapping : Visualizing Technological Distance. *Journal of the Association for Information Science and Technology* **65**, 2432–2443 (2014).

20. L. Leydesdorff, D. Kushnir, I. Rafols, Interactive Overlay Maps for US Patent (USPTO) Data Based on International Patent Classification (IPC). *Scientometrics* **98**, 1583–1599 (2014).

21. B. Verspagen, Measuring Intersectoral Technology Spillovers: Estimates from the European and US Patent Office Databases. *Economic Systems Research* **9**, 47–65 (Mar. 1, 1997).

22. F. Neffke, M. Henning, Skill Relatedness and Firm Diversification. *Strategic Management Journal* **34**, 297–316 (Mar. 1, 2013).

23. G. Bottazzi, D. Pirino, "Measuring Industry Relatedness and Corporate Coherence", Working Paper 2010/10 (LEM, 2010).

24. B. Leten, R. Belderbos, B. Van Looy, Technological Diversification, Coherence, and Performance of Firms. *Journal of Product Innovation Management* **24**, 567–579 (Nov. 1, 2007).

25. S. Breschi, F. Lissoni, F. Malerba, Knowledge-Relatedness in Firm Technological Diversification. *Research Policy* **32**, 69–87 (Jan. 2003).

26. D. L. Rigby, Technological Relatedness and Knowledge Space: Entry and Exit of US Cities from Patent Classes. *Regional Studies* **49**, 1922–1937 (2015).

27. C. Castaldi, K. Frenken, B. Los, Related Variety, Unrelated Variety and Technological Breakthroughs: An Analysis of US State-Level Patenting. *Regional Studies* **49**, 767–781 (May 4, 2015).

28. F. Neffke, M. Henning, R. Boschma, How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Economic Geography* **87**, 237–265 (2011).

29. K. Frenken, F. Van Oort, T. Verburg, Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies* **41**, 685–697 (July 2007).

30. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The Product Space Conditions the Development of Nations. *Science* **317**, 482–7 (July 2007).

31. A. Hatchuel, B. Weil, presented at the DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm.

32. A. Hatchuel, P. Le Masson, B. Weil, presented at the DS 32: Proceedings of DESIGN 2004, the 8th International Design Conference, Dubrovnik, Croatia.

33. S. Wuchty, B. F. Jones, B. Uzzi, The Increasing Dominance of Teams in Production of Knowledge. *Science* **316**, 1036–1039 (May 18, 2007).

34. D. Malva, Antonio, M. Riccaboni, "(Un)Conventional Combinations: At the Origins of Breakthrough Inventions", SSRN Scholarly Paper ID 2610562 (Social Science Research Network, Rochester, NY, Oct. 1, 2014).

35. A. Phene, K. Fladmoe-Lindquist, L. Marsh, Breakthrough Innovations in the U.S. Biotechnology Industry: The Effects of Technological Space and Geographic Origin. *Strategic Management Journal* **27**, 369–388 (Apr. 1, 2006).

36. J. Chan, S. P. Dow, C. D. Schunn, Do the Best Design Ideas (Really) Come from Conceptually Distant Sources of Inspiration? *Design Studies* **36**, 31–58 (Jan. 2015).

37. O. Shai, Y. Reich, Infused Design. I. Theory. *Research in Engineering Design* **15**, 93–107 (2004).

38. I. Tseng, J. Moss, J. Cagan, K. Kotovsky, The Role of Timing and Analogical Similarity in the Stimulation of Idea Generation in Design. *Design Studies* **29**, 203–221 (2008).

39. D. Gentner, A. B. Markman, Structure Mapping in Analogy and Similarity. *American psychologist* **52**, 45 (1997).

40. J. O. Wilson, D. Rosen, B. A. Nelson, J. Yen, The Effects of Biological Examples in Idea Generation. *Design Studies* **31**, 169–186 (2010).

41. K. Fu, J. Cagan, K. Kotovsky, K. Wood, Discovering Structure in Design Databases Through Functional and Surface Based Mapping. *Journal of Mechanical Design* **135**, 031006–031006 (Feb. 20, 2013).

42. K. Fu, J. Murphy, M. Yang, K. Otto, D. Jensen, K. Wood, Design-by-Analogy: Experimental Evaluation of a Functional Analogy Search Methodology for Concept Generation Improvement. *Research in Engineering Design* **26**, 77–95 (2015).

43. J. Chan, K. Fu, C. Schunn, J. Cagan, K. Wood, K. Kotovsky, On the Benefits and Pitfalls of Analogies for Innovative Design: Ideation Performance Based on Analogical Distance, Commonness, and Modality of Examples. *Journal of Mechanical Design* **133**, 081004–081004 (Aug. 1, 2011).

44. L. Fleming, D. M. Waguespack, Brokerage, Boundary Spanning, and Leadership in Open Innovation Communities. *Organization Science* **18**, 165–180 (Apr. 2007).

45. A. Schoen, L. Villard, P. Laurens, J.-p. Cointet, G. Heimeriks, presented at the Science & Technology Inidicators.

46. A. B. Jaffe, M. Trajtenberg, *Patents, Citations, and Innovations: A Window on the Knowledge Economy* (MIT Press, 2002), 502 pp.

47. A. B. Jaffe, G. de Rassenfosse, "Patent Citation Data in Social Science Research: Overview and Best Practices", Working Paper 21868 (National Bureau of Economic Research, Jan. 2016).

48. S. Valverde, R. V. Solé, M. A. Bedau, N. Packard, Topology and Evolution of Technology Innovation Networks. *Physical Review E* **76**, 056118 (Nov. 28, 2007).

49. B. H. Hall, R. H. Ziedonis, The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979-1995. *The RAND Journal of Economics* **32**, 101–128 (Apr. 1, 2001).

50. K. Börner, N. Contractor, H. J. Falk-Krzesinski, S. M. Fiore, K. L. Hall, J. Keyton, B. Spring, D. Stokols, W. Trochim, B. Uzzi, A Multi-Level Systems Perspective for the Science of Team Science. *Science Translational Medicine* **2**, 49cm24–49cm24 (Sept. 15, 2010).

51. S. M. Fiore, Interdisciplinarity as Teamwork: How the Science of Teams Can Inform Team Science. *Small Group Research* **39**, 251–277 (June 1, 2008).

52. H. D. White, B. Wellman, N. Nazer, Does Citation Reflect Social Structure?: Longitudinal Evidence from the "Globenet" Interdisciplinary Research Group. *Journal of the American Society for Information Science and Technology* **55**, 111–126 (Jan. 15, 2004).

53. R. Conti, A. Gambardella, M. Mariani, Learning to Be Edison: Inventors, Organizations, and Breakthrough Inventions. *Organization Science* **25**, 833–849 (Dec. 23, 2013).

54. B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A Probabilistic Programming Language. *Journal of Statistical Software* (2016 (in press)).

55. D. Brockmann, L. Hufnagel, T. Geisel, The Scaling Laws of Human Travel. *Nature* **439**, 462–465 (Jan. 26, 2006).

56. P. Deville, C. Song, N. Eagle, V. D. Blondel, A.-L. Barabási, D. Wang, Scaling Identity Connects Human Mobility and Social Interactions. *Proceedings of the National Academy of Sciences* **113**, 7047–7052 (June 28, 2016).

57. M. Trajtenberg, A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The RAND Journal of Economics* **21**, 172–187 (1990).

58. M. B. Albert, D. Avery, F. Narin, P. McAllister, Direct Validation of Citation Counts as Indicators of Industrially Important Patents. *Research Policy* **20**, 251–259 (1991).

59. D. Harhoff, F. Narin, F. M. Scherer, K. Vopel, Citation Frequency and the Value of Patented Inventions. *Review of Economics and Statistics* **81**, 511–515 (Aug. 1, 1999).

60. B. Hall, A. Jaffe, M. Trajtenberg, "Market Value and Patent Citations: A First Look", NBER Working Paper 7741 (National Bureau of Economic Research, Inc, June 2000).

61. A. Gambardella, D. Harhoff, B. Verspagen, The Value of European Patents. *European Management Review* **5**, 69–84 (2008).

62. B. Nooteboom, *Learning and Innovation in Organizations and Economies* (OUP Oxford, 2000).

63. P. L. Masson, A. Hatchuel, O. Kokshagina, B. Weil, Designing Techniques for Systemic Impact: Lessons from C-K Theory and Matroid Structures. *Research in Engineering Design* **28**, 275–298 (July 1, 2017).

Appendix

Data

The domain of the invention (classification)

For a patent to be granted, the invention described in the application must be considered sufficiently novel and non-obvious by a patent examiner who is an expert trained in the subject matter. To ensure that a patent is examined by an examiner who is actually an expert in the domain, it is necessary to match the patent with a domain, and then the patent can be assigned to the correct patent examiners. Similarly, in order to ensure that the patent describes an invention that is novel, the patent examiner should be able to access a set of patents within the domain of the patent, so as to compare them. For this purpose it is also useful to have each patent matched with a domain, for ease of finding past patents that are relevant and for being found as relevant in future examination of other patents. For these purposes, patent examiners classify patents into technology classifications systems. Individual patents' classifications can be and frequently are updated throughout the examination process, as the patent examiners have incentive to classify the patent as accurately as possible and to update the classification if it is inaccurate. Such classification updating can also occur once the patent is awarded: if a new domain of technology arises (e.g. 3D printing), a new technology class is eventually added to the classification system (*1*). At this point, the USPTO goes back through all previously granted patents to re-classify those that belong to the newly-recognized technology domain. The reason for this re-classification is so that these older patents will still be found during future examinations of patents in the newly-recognized domains. Patent classification is thus a domain-expert-curated, repeatedly-updated assessment of what domains best describe millions of inventions.

Patent offices around multiple technology classification systems. We used patent class data from the International Patent Classification system (IPC), curated by the World Intellectual Property Organization. Like most technology classification systems, the IPC is hierarchical: There are sections (e.g. 'B: Performing Operations; Transporting') divided into classes ('B64: Aircraft; Aviation; Cosmonautics') divided into sub-classes ('B64C: Aeroplanes; Helicopters'). We used the sub-classes at the "4-digit" to represent the technology domains, of which there were 629. We took each patent's main classification at the 4-digit level to be the technology domain of that patent.

The hierarchical classification system goes to further levels of detail, and it is conceivable to go deeper into the hierarchy to perform the present analysis with a higher-resolution description of technology domains. However, the predictive model of inventors' explorations involved retaining data not just on each domain that an inventor entered, but also all the domains that they did not enter. Keeping track of over 600 domains for each such move pushed the limits of computational tractability, and keeping track of thousands or tens of thousands of domains for each move would go beyond what is computationally feasible at present.

Citations to other patents

Patents typically contain citations to other patents, and the purpose of these citations to highlight the limits of what the patent can claim as novel intellectual property. For example, if a patent describes a telescoping fishing rod, it may cite a patent for a non-telescoping fishing rod and another patent using a telescoping mechanism in another context. This would make clear that the patent does not have claims to the ideas of fishing rods or telescoping mechanisms, but to specifically the telescoping fishing rod. Patent citations have been studied for decades as a signal how inventions build on other inventions (*2*). There remains questions of how much individual inventors are actually aware of these other inventions during their own invention process, as the most citations are not made by the inventors themselves, but by lawyers and patent examiners (*3*, *4*). At the very least, it is clear a citation between two patents indicates that they describe inventions that are related in some way. We use citations as a signal of relatedness between technology domains, though that signal must be extracted from spurious factors (described below).

The number of citations between domains is directed: the number of citations from “semiconductors” to “photography” could be different from the number of citations from “photography” to “semiconductors”. However, in practice the two directions were very strongly correlated (Pearson’s r: .9959). Here we used the number of citations each domain received from other domains. For example, Fig. 1 describes the number of citations that “semiconductors” received from “static info storage”, “photography”, and “hydraulics”. For the present analysis of inventors’ movements across domains , we calculated the R to unentered domains by using the citations received from the domains in the inventor’s existing portfolio. All results using citations in the reverse direction, citations made by the domain, are virtually identical.

Who invented the invention (author names)

Patent’s inventors are listed simply as a name and an address; there is no ID number for each inventor that identifies them across all their patents. Additionally, the address is only a city (no street address), and the name is not a complete legal name: the same inventor could be listed as “Joe Smith” on one patent and “Joseph C. Smith” in another patent. The lack of precision in labeling patents with their inventors’ identities has lead to research on how to disambiguate inventors’ names across patents. We used the name disambiguation data provided by (*5*), wherein names were disambiguated with a probabilistic model. This model identified inventor names as being more similar (and thus more likely to be labeled as referring to the same person) based on 1. the lexicographical distance between the first, middle and last names, 2. the physical distance between the names’ associated geographic addresses, 3. the number of shared co-authors, 4. whether the names both are listed as authors on patents that are assigned to the same organization (e.g. IBM or MIT), and 5. the number of technology classes the names’ patents share.

The fact that the name disambiguation algorithm identified individual inventors using information about technology classes likely affects the present analysis of inventors’ explorations across technology domains. If two names (“Jen Yu” and “Jennifer T. Yu”) are written on patents that are classified into different technology classes, the two names will likely be identified as different inventors. Sometimes these two names will, in fact, refer to the same inventor, who has simply patented in multiple domains. The name disambiguation algorithm thus introduces a bias: explorer inventors will sometimes be identified as multiple individual inventors who do not explore. Therefore, exploration will appear less frequent than it actually is.

The USPTO has recently employed in-house an inventor name disambiguation scheme that is more sophisticated (and likely accurate) than the scheme used in the present study (*6*). This algorithm employs an advanced data matching system to quickly find optimal name disambiguation, and along the way it uses multiple kinds of data. One of those data points, unfortunately, is problematic for our purposes. The algorithm of (*6*) uses natural language processing of patents' titles to measure if the patents of two inventor names are about similar subjects, and thus the two names are likely to refer to the same inventor. Assuming that the text-based analysis of patent similarity and the citation-based analysis of domain relatedness have any agreement, then using this name disambiguation scheme would thus build into the data the very phenomenon we are examining: that inventors are more likely to explore related domains. We have thus not used this data, so as to not tautologically assume the hypothesis of how inventors explore. It is likely that future work could use this name disambiguation algorithm (and its descendents) to achieve more accurate prediction of inventors' explorations, but the interpretation of relatedness would be more epistemically fraught.

Power laws indicate inventors actually explore

One simple way we could be misinterpreting the data is that inventors do not explore new domains at all, but the apparent "movements" are an artifact of the classification system. Imagine an inventor whose patents are all in a single, coherent field, but the classification system does not crisply capture that field; instead, patents in that field are randomly assigned to one of two domains that together approximate the topic. The inventor's patent record would show them inventing in one domain, then "moving" into another. We might even expect relatedness to predict the next domain: "related" pairs of domains would presumably be more likely to share a coherent field. However, this scenario wouldn't explain the rate of exploring new domains (Fig. 3): random patent classification would yield an exponential drop-off in new domain entries with additional patents, and the empirical shape is better-described by a power law. It is possible to create a distribution that looks like a power law by superimposing multiple exponential distributions with different exponents (*7*), which could be the case if some of the "exploring" inventors' true fields were spread across 2 domains, and others 3 domains, and others 4 domains, etc. However, achieving the power law appearance would require mixing the exponential distributions in equal quantities, which would require that there were roughly equal number of true fields that were split across 2 domains as 3 domains as 4 domains, etc. This seems less likely than the alternative: expert-curated patent classification systems are imperfect but fairly accurate, and inventors explore new domains with dynamics like a power law, just as occurs in other aspects of human behavior (*8*).

Prediction with a Discrete Naive Classifier

The predictive model was a kind of discrete naive classifier. It was discrete, in that the data was discretized and no smoothing was applied. It was naive, in that it included no joint conditional probabilities (e.g. $p(\text{entry}|R, \text{popularity})$).

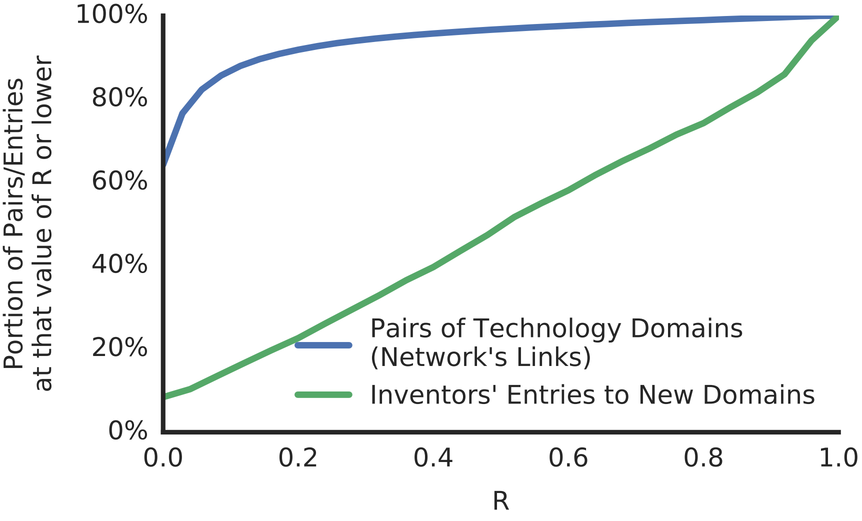


Figure A1. The portion of pairs of domains with low *R* was very high, but inventors' entries were more evenly distributed across the values of *R*. The *R* associated with each entry (green line) is calculated using patent data from the years before that entry. The *R* associated with each pair of domains (blue line) is calculated using patent data from 2010. The differences in date biases the two lines to be closer together (the *R* values of the blue line to be lower and of the green line to be higher), but they are still clearly distinct.

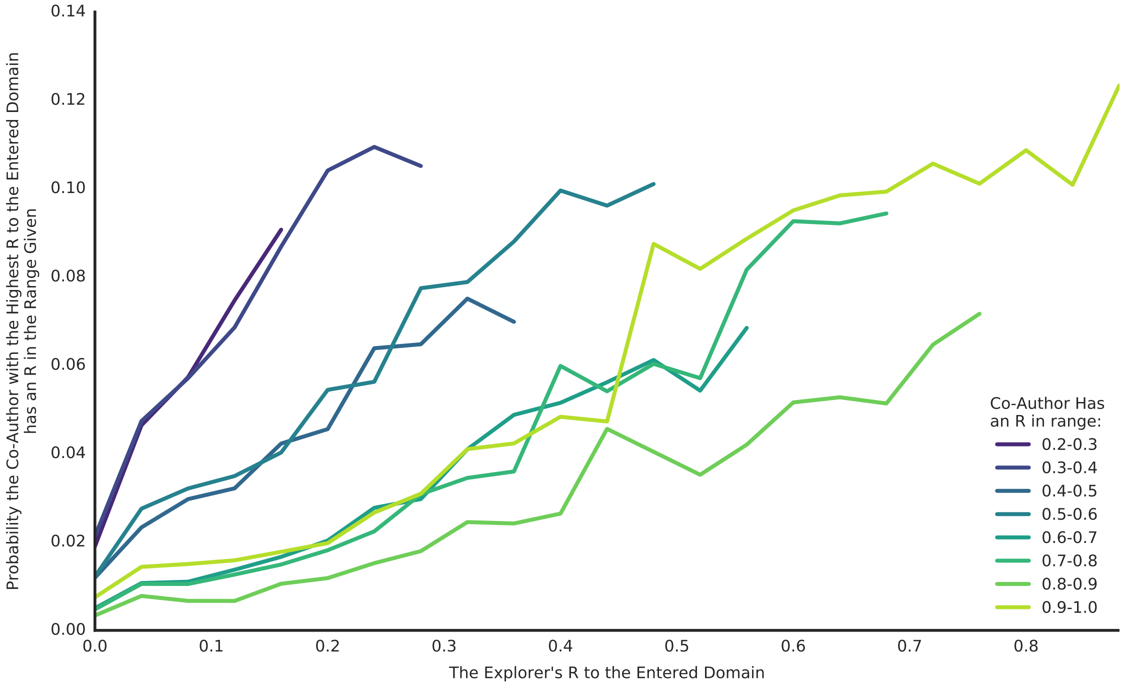


Figure A2. Explorers' first patent in an entered domain was more likely to have co-authors with higher *R* to the domain if the explorer had a higher *R* themself. As Fig. 8, but the co-author's *R* is within a specific range (legend).

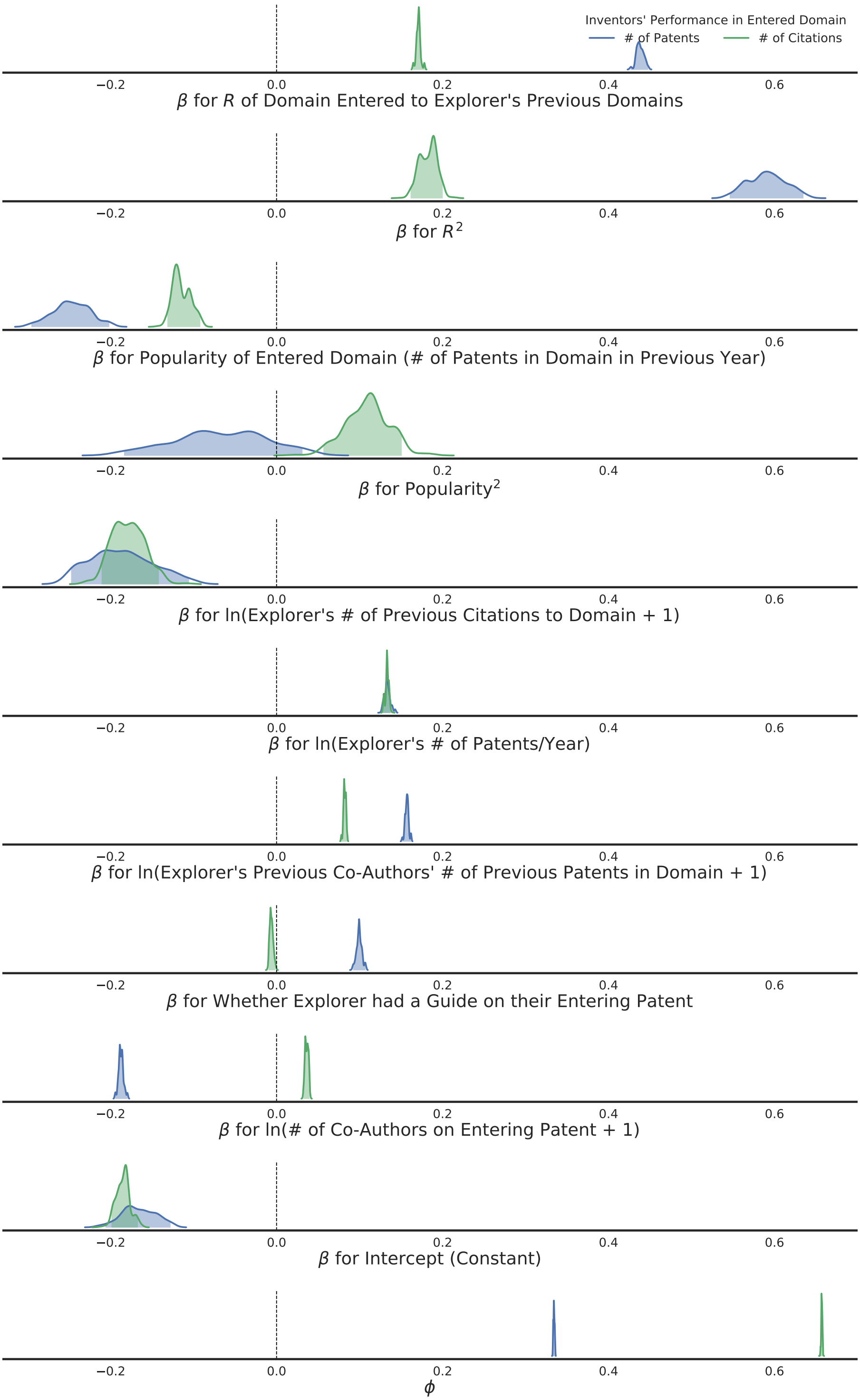


Figure A3. Parameter values for models of inventors' future number of patents (blue) and citations (green) in an entered domain. Lines: kernel density estimates of the posterior distribution of each parameter's values. Shading: parameters' 95% credible interval.

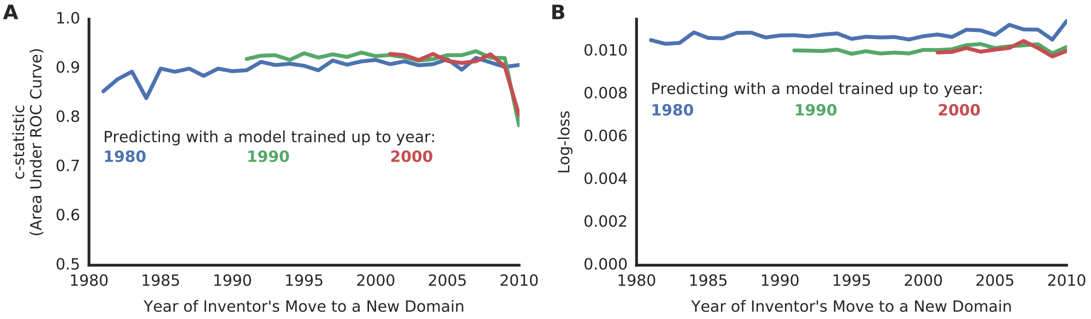


Figure A4. Multiple measures of the predictive model's power show persistently accurate prediction on long time horizons. Predictive models were created using data from 1976 up to 1980 (blue), 1990 (green) and 2000 (red). These models were then used to predict explorers' movements in subsequent years, after the time period included in the model training. A) The c-statistic of the models' predictions (area under the receiver operating characteristic curve). B) The logarithmic loss of the models' predictions.

The relationships between the variables shown in Fig. 4B,C *are* joint conditional probabilities, but they were not used for prediction. Conditional probabilities could undoubtedly be incorporated to achieve higher prediction, but the increased dimensionality of the model would create challenges with sparse data (e.g. an observation of $(R = 1, \text{popularity} = 58.2, \text{co-authors} = True, \text{citations} = False)$ having no historical precedent, even though there was a previous observation of $(R = 1, \text{popularity} = 58.4, \text{co-authors} = True, \text{citations} = False)$). Addressing higher dimensionality would require smoothing the data, which essentially introduces a prior. While discretizing the data is also a form of unsophisticated prior, we did not want to assume that data had any particular functional form (beyond that it could be approximated by a histogram). Thus, we kept the predictive model to a discrete, naive classifier.

Alternative Definition of Relatedness: Inventors Co-Occurrence and Co-Classification

This study has focused on measuring the relatedness between technology domains by using the citation behavior of patents. However, there are other ways to use patent data to measure how much domains interact, and thus how they may be related. Two common techniques are Co-Occurrence and Co-Classification.

Inventor Co-Occurrence

A common technique to assess if there is a latent connection between two domains (be they technology domains, or product categories, etc.) is to measure how often the two domains occur simultaneously in the same portfolio of some entity (be that an inventor, an organization, a country, etc.). As more concrete examples, Co-Occurrence metrics have been used to quantify the connections between two products by observing how often a country that exports one product exports the other (*9*), and Co-Occurrence metrics have also been used to quantify the connections between two technology domains by observing how often a firm that produces patents in one domain also produces patents in the other (*10*). Clearly, when describing inventor behavior a relevant Co-Occurrence metric to use to measure the relatedness of two domains is how often an inventor who patents in one domain also patents in the other domain: Inventor Co-Occurrence.

Inventor Co-Occurrence, like citation behavior, is affected by phenomena that are not technology relatedness, like the number of domains each inventor has entered and the popularity of each domain. As was done with citation behavior, we calculated R for Inventor Co-Occurrence by comparison to the expected number of co-occurrences by chance. This expectation was created using the methods described in (*10*). The quantifications of domains' relatedness using normalized Inventor Co-Occurrence and using normalized citation counts are very correlated (*10*), and so measuring R with either yielded qualitatively similar results in prediction and performance (Figs. A6, A5).

It is worth noting that Co-Occurrence has a theoretical difficulty: it does not give a hint of a mechanism for inventors' explorations, because the measure *is* inventors' explorations. Measuring Co-Occurrence with data from 1976 to 1980 is effectively summarizing the paths of inventors' explorations from 1976 to 1980. That Co-Occurrence can be used to then predict data from 1981 may seem tautological, but it does in fact indicate something: it is evidence that the paths of exploration across domains are relatively stable, so that the same paths taken in the past are close to those taken in the immediate future. However, once we have identified those paths, we do not know why they are where they are. To learn this we need additional data. There are many hypotheses for why paths are where they are, which can be addressed with different kinds of data. One hypothesis is that a commonly-used path between two domains arises from the two domains requiring a similar knowledge base, which may be reflected in citations. No matter what the mechanisms actually are for creating these paths, Co-Occurrence cannot see them. Being blind to the underlying mechanisms is particularly relevant if those mechanisms lead to a change, like if a new set of knowledge links two domains that had not Co-Occurred before (e.g. semiconductors and photography becoming related in the 1990s). In practice such changes are rare and slow (as evidenced by the generally strong predictive

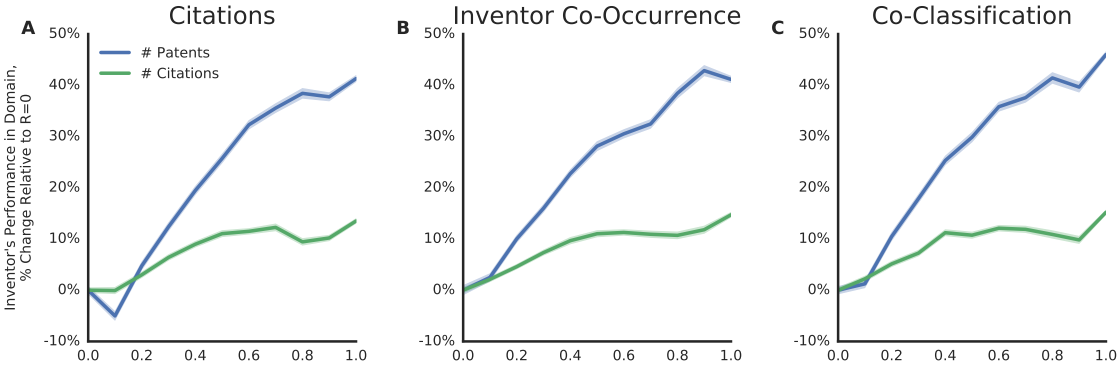


Figure A5. The relationships between *R* and explorers' performance were robust to different measures of technology relatedness.

power of Co-Occurrence), but identifying and explaining these changes are an opportunity for future study.

Co-Classification

When a patent is classified it is assigned a main or primary class, which is the class used for all other analyses in this study. However, some patents were assigned one or more secondary classes. Secondary classifications indicated technology domains that were not where the primary inventive contribution of the patent lay, but were still components or aspects of the invention described (typically reported in claims other than the first one). Only about 17% of patents were assigned one or more secondary classes at the 4-digit level of the IPC classification system. Still, it is possible to use instances of multiple classification to measure how often two technology domains appear on the same patent: Co-Classification. Co-Classification is mathematically the same as Co-Occurrence (the entity with a portfolio of technology domains is just a patent, not a person), and so we measured Co-Classification using the same normalization process. Normalized Co-Classification is also correlated with normalized citation counts (*10*), and so measuring *R* using Co-Classification yielded qualitatively similar results to measuring *R* with citations (Figs. A5,A6).

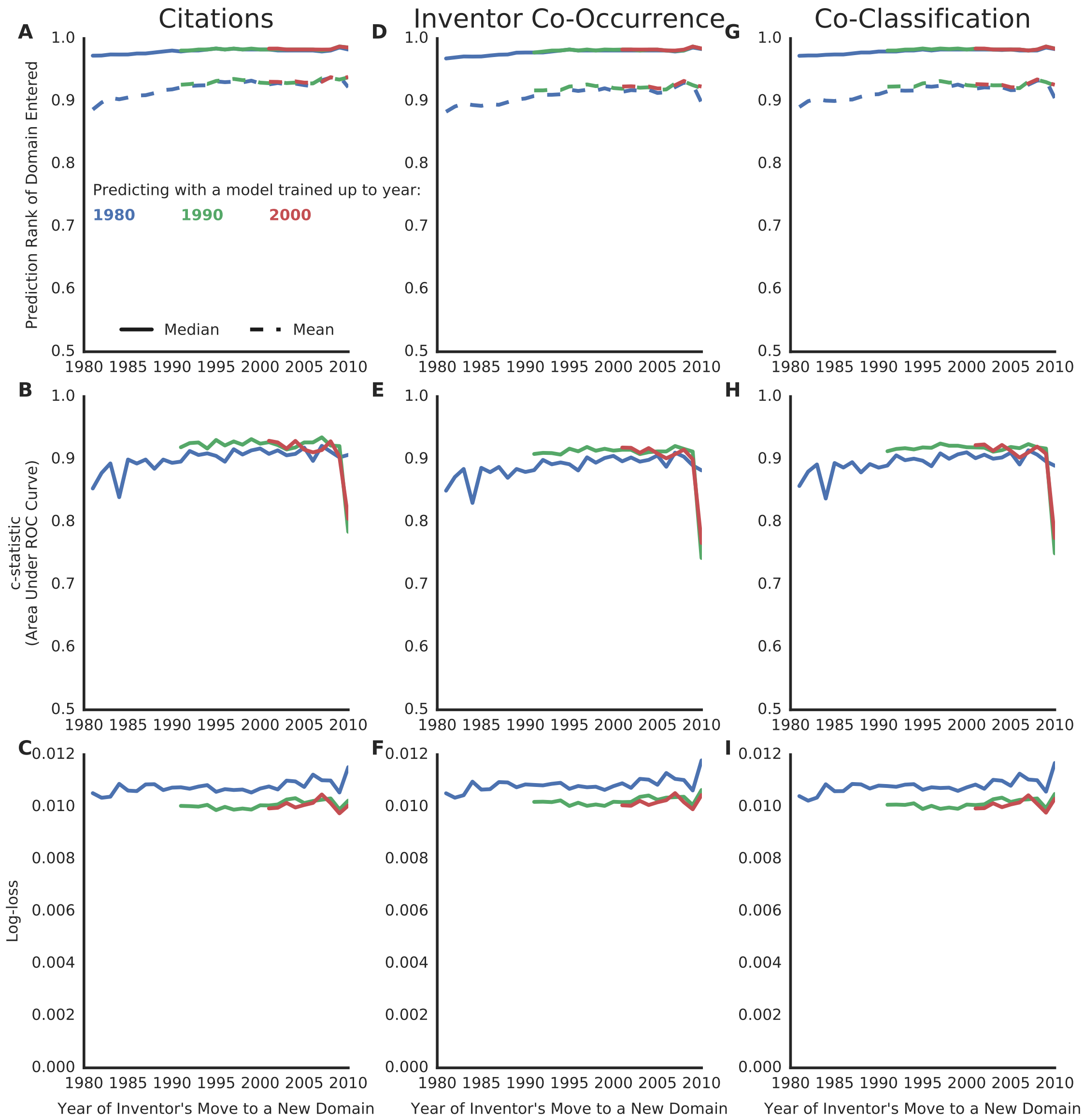


Figure A6. The predictability of explorers' future moves was maintained when using different measures of domains' interactions to quantify technology relatedness. The predictive power of three different models, measuring domains' relatedness through three different kinds of interactions: their number of citations (Citations: A-C), how often an inventor's portfolio has patents in both domains (Inventor Co-Occurrence: D-F), how often a patent is classified in both domains simultaneously (Co-Classification: G-I).

References

1. F. Lafond, D. Kim, Long-Run Dynamics of the U.S. Patent Classification System. arXiv: 1703.02104 (q-fin) (Feb. 27, 2017).
2. A. B. Jaffe, G. de Rassenfosse, “Patent Citation Data in Social Science Research: Overview and Best Practices”, Working Paper 21868 (National Bureau of Economic Research, Jan. 2016).
3. J. Alcácer, M. Gittelman, Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *Review of Economics and Statistics* **88**, 774–779 (Nov. 1, 2006).
4. P. Criscuolo, B. Verspagen, Does It Matter Where Patent Citations Come from? Inventor vs. Examiner Citations in European Patents. *Research Policy*, Special Section Knowledge Dynamics out of Balance: Knowledge Biased, Skewed and Unmatched **37**, 1892–1908 (Dec. 2008).
5. G.-C. Li, R. Lai, A. D’Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, L. Fleming, Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975-2010). *Research Policy* **43**, 941–955 (July 2014).
6. N. Monath, A. McCallum, “Discriminative Hierarchical Coreference for Inventor Disambiguation”, Alexandria, VA, Sept. 24, 2015.
7. J. Chu-Shore, M. B. Westover, M. T. Bianchi, Power Law versus Exponential State Transition Dynamics: Application to Sleep-Wake Architecture. *PLoS ONE* **5**, e14204 (Dec. 2, 2010).
8. P. Deville, C. Song, N. Eagle, V. D. Blondel, A.-L. Barabási, D. Wang, Scaling Identity Connects Human Mobility and Social Interactions. *Proceedings of the National Academy of Sciences* **113**, 7047–7052 (June 28, 2016).
9. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The Product Space Conditions the Development of Nations. *Science* **317**, 482–7 (July 2007).
10. J. Alstott, G. Triulzi, B. Yan, J. Luo, Mapping Technology Space by Normalizing Patent Networks. *Scientometrics* **110** (2017).