SOCIAL NETWORKS

The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot

Satyam Mukherjee,[1,2] Daniel M. Romero,[1,2,3] Ben Jones,[1,4] Brian Uzzi[1,2]*

Scientists and inventors can draw on an ever-expanding literature for the building blocks of tomorrow's ideas, yet little is known about how combinations of past work are related to future discoveries. Our analysis parameterizes the age distribution of a work's references and revealed three links between the age of prior knowledge and hit papers and patents. First, works that cite literature with a low mean age and high age variance are in a citation "hotspot"; these works double their likelihood of being in the top 5% or better of citations. Second, the hotspot is nearly universal in all branches of science and technology and is increasingly predictive of a work's future citation impact. Third, a scientist or inventor is significantly more likely to write a paper in the hotspot when they are coauthoring than whey they are working alone. Our findings are based on all 28,426,345 scientific papers in the Web of Science, 1945–2013, and all 5,382,833 U.S. patents, 1950–2010, and reveal new antecedents of high-impact science and the link between prior literature and tomorrow's breakthrough ideas.

INTRODUCTION

Scientists and inventors can combine information from an ever-expanding knowledge base dispersed across documents, experiments, and data (*1*). The Web of Science (WOS) contains 28.4 million publications, including more than 1.5 million new articles published in 2014 alone, sextupling the 1970 rate. The U.S. Patent and Trademark Office (USPTO) issued 287,831 patents in 2013, quadrupling the 1970 rate. Although more knowledge enables more novel ideas to be combined (*2*, *3*), scientists and inventors have limited time to search through the expanding base (*4*, *5*). As the rate of knowledge expansion grows, but the time to search for new knowledge remains fixed, scientists and inventors search a smaller fraction of the available knowledge (*6*–*8*). These trade-offs between available knowledge and search costs make understanding where to search for the most valuable past information important to new knowledge advances (*2*, *4*, *7*). However, little research exists on where in the store of knowledge to find the best combinations of past information (*2*).

Consider the following case. Imagine you are in the library of Alexandria in 48 BCE. At the time, the library of Alexandria is the largest store of scientific knowledge on the planet and growing rapidly. Every document related to science, philosophy, or religion written in Egypt is copied and put into the library, as is every document found on every caravan or boat that lands on Egyptian shores. In 48 BCE, the library catches fire. You now have limited time to search for knowledge you think is going to be most valuable for creating the next set of important ideas in your field. How do you search the store of knowledge? Do you gather up the most recent documents under the assumption that they offer a summary statistic of the best of past knowledge? Do you collect the oldest papers that have stood the test of time? Do you look for the papers that were read by the most experts in your field? Do you sample papers at random?

Theories of knowledge development emphasize the importance of past information in the formulation of new ideas (*2*, *9*–*12*) but offer different answers to the questions about where to search for the most fruitful information. One school of thought argues that older work, benefiting from the test of time, is most likely to provide the building blocks of new work, an idea reflected in Isaac Newton's famous remark, "If I have seen further than others, it is by standing upon the shoulders of giants" (*10*). By contrast, Robert Merton's births of time theory (*11*) suggests that recent information drives breakthrough ideas (*11*, *12*). Consistent with Merton's formulation, many information retrieval systems search for recent information first. Cognitively, people tend to retrieve the most recent information first (*13*). Search engines typically return results according to either recency or popularity, both of which correlate with the age of the information (*14*).

Diverging arguments and little empirical study on the information search question have meant that the link between the age of information referenced in a work and a work's impact remains an open question. Knowing whether old, new, randomly sampled, or popular information is associated with the creation of novel combinations can help provide insight into where the richest combinations of past knowledge are located. To address these questions, we studied modern science and invention to identify the empirical patterns linking the age of information cited in a paper or patent and the paper's or patent's subsequent impact.

RESULTS

We investigated two large domains of knowledge: all 28,426,345 papers in the WOS, 1945–2013, and all 5,382,833 patents published in the U.S. patent office database, 1950–2010. In both domains, the references cited in a work identify the age of the past knowledge it builds upon (*2*, *3*, *5*). To quantify the age of information referenced in a work, we computed the age distribution of its cited references (*6*, *9*). This distribution, denoted D, contains the age differences between a work's publication year and the publication years of its references. Two properties of D are its mean (D_μ) and coefficient of variation (COV) (D_θ), which we computed for each scientific paper in the WOS and for each U.S. patent. Figure S1 (A and B) shows the empirical distributions of D_μ and D_θ for all papers in the WOS in 1995. Figures S2 and S3 (A to C) present null models of referencing behavior and indicate that the observed distributions of D_μ and D_θ are not explained by chance. Table S2 shows the measurements of D_μ and D_θ for four example papers.

A common measure of the impact of a scientific paper or patent is the number of citations it receives (*2*, *3*, *12*, *15*). We defined a work as high impact ($H = 1$) if a work is in the top 5th percentile of cited works

[1]Northwestern University, Evanston, IL 60208, USA. [2]Northwestern Institute on Complex Systems and Data Science, Evanston, IL 60208, USA. [3]University of Michigan, Ann Arbor, MI 48109, USA. [4]National Bureau of Economic Research, Cambridge, MA 02138, USA.
*Corresponding author. Email: uzzi@northwestern.edu

in its scientific or technological subfield based on the citations it accumulated in the first 8 years after publication (*2*, *15*, *16*), and low impact otherwise ($H = 0$). As robustness checks on our measure, we also measured impact as being in the top 1, 10, and 25% of the citation distribution, the log of the number of citations a paper accumulates after 8 years of publication and over its lifetime (*12*, *17*), and a paper's PageRank of citations, that is, papers with a possibly low number of absolute citations but with relatively many cites from hit papers (*18*). Below, we present the results for papers in the top 5% of the citation distribution. In Materials and Methods, we provide detail on measurements. Robustness checks using other measures of H (impact) and D (age of knowledge) are presented in the Supplementary Materials (tables S1 and S3 to S9). All measures produced similar results.

The knowledge hotspot and scientific impact

Figure 1 is a heat plot of the relationship between D_μ, D_θ, and H. Each point in the plot represents the D_μ and D_θ values of papers published in 1995 ($N = 546{,}912$), and the intensity of the color represents a paper's probability of high impact. The plot's vertical and horizontal lines represent the median and mean population-level values of D_μ and D_θ. Three main findings are demonstrated. First, amid all the distributions of the ages of information, one type of age distribution is especially associated with high impact. Papers in this "hotspot" have a low D_μ and high D_θ and are 2.2 times more likely, on average, to receive citations at a level of 5% or better in their field. Figure 2 further characterizes the knowledge space of papers in the hotspot with respect to time. Papers in the hotspot reference recent ideas in the literature (low average D_μ = 6.05, SD = 1.74) and ideas of a relatively wide variation of age (high D_θ = 1.0, SD = 0.23), as revealed by the tail of this distribution, which reaches well into the past at a progressively decreasing rate. Second, papers that center their references on new knowledge—low D_μ and low D_θ—have a surprisingly low rate of impact that rarely exceeds what is expected by chance. This suggests that the conventional bias toward heavily citing recent work (*6*) is valuable only when mixed with a high D_θ. Third, papers that reference prior work centered on older knowledge—that is, papers with high D_μ and low D_θ (27% of all papers)—are notable in that they have an H that is half the rate expected by chance.

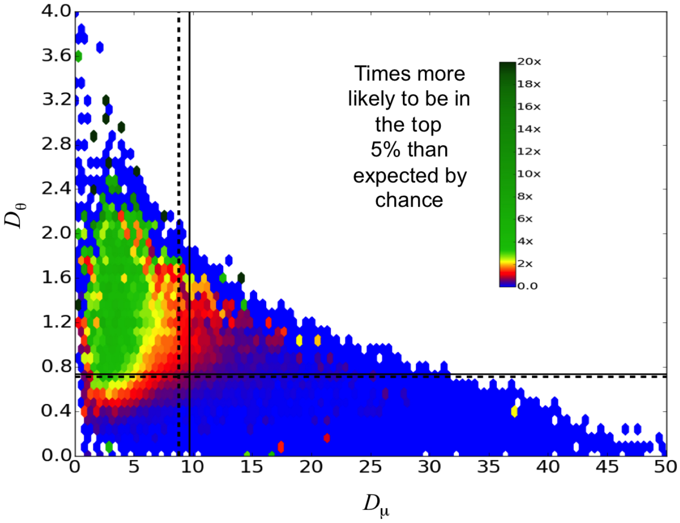

Fig. 1. Knowledge hotspot predicts high-impact science. Papers in the hotspot are, on average, more than two times as likely to be hits than the background rate (data shown are for the year 1995, $N = 546{,}912$ papers). The hotspot is the overrepresented concentration of "hit" papers shown in green that cite prior knowledge with a low mean age, D_μ, and a high age COV, D_θ. Notably, 75% of papers are outside the hotspot, and their likelihood of being a hit is no greater than expected by chance. Solid lines and dotted lines are population means and medians of D_μ and D_θ. The background rate is the likelihood of a paper chosen at random being in the top 5% of citations for papers in that field.

Figure 3 demonstrates that the relationship between the hotspot and level of impact has been remarkably robust across time. Pooling all WOS papers on a year-by-year basis from 1950 to 2005, we find that the information hotspot has invariantly been strongly related to high-impact work for all of modern science. Scientific papers in the hotspot consistently double their chances of being a hit. Further, we observe a growing trend of a paper being a hit when it is in the hotspot. By contrast, papers outside the hotspot have risen and fallen in their relationship to impact but generally remain relatively low impact, with no other mix of D_μ and D_θ exceeding the 5% background rate expected by chance. This empirical regularity suggests a fundamental ordering that may characterize the relationship between the age of information referenced in a scientific paper and extraordinary scientific impact.

Figure 4 disaggregates the data in the WOS on a field-by-field basis, revealing a marked similarity across the branches of science with regard to the main findings. The WOS lists 171 subfields in science and engineering, 54 subfields in social sciences, and 27 subfields in arts and humanities. Using the classification for the four types of information search shown in Fig. 1, we computed the fraction of WOS fields for which papers in the hotspot are associated with the highest citation impact. Figure 4 demonstrates that, at the beginning of the postwar era of science, about 60% of fields displayed the "hotspot-hit link" (green bar). By the 2000s, the hotspot overrepresents hit papers in almost 90% of the fields. Thus, despite the large differences between scientific fields in terms of theory, methods, data, and culture, the hotspot dominates the sciences.

To test these patterns in the data while controlling for other variables, we ran fixed-effects regressions to predict the citation impact of individual papers. Fixed-effects regressions allows us to control in a nonparametric and flexible manner for numerous features of each paper, including the predictive capacity of each (i) field, (ii) publication year, (iii) number of references made, and (iv) number of authors. In addition, we control for the degree to which a work references (v) prior work from multiple/interdisciplinary fields, (vi) highly cited papers (*19*), and (vii) conventional and/or novel pairings of prior ideas (*2*). (See Methods for the fixed-effects regression model, the variable construction details, and the related approach used for patents.)

Table 1 shows that the regression models indicate three important relationships between H, D_μ, and D_θ. First, the knowledge hotspot is strongly related to citation impact net of control variables for time, all 254 scientific fields, and paper-level characteristics. The large drops in the Bayesian information criterion (BIC) statistics when D_μ and D_θ are added to the control variable regressions indicate the strong explanatory power of D_μ and D_θ (Materials and Methods). Similarly, standardizing the regression coefficients indicates that D_μ and D_θ have large substantive effects on citation impact relative to other predictors of citation impact (table S10). Second, Fig. 5 reveals the intricate joint behavior of D_μ and D_θ in relation to hit papers. Papers with a high mean age of references (D_μ) are always associated with a low probability of being a hit irrespective of the variation in the age of references (D_θ). Conversely, a low mean age of references (D_μ) is associated with being a hit only when the age variance of references (D_θ) is high. Papers with a low D_μ and low D_θ have surprisingly no greater likelihood of being a hit than expected by chance. Third, the above findings are robust to diverse measures of a hit. Tables S2 to S8 show that the above results are replicated when H is measured

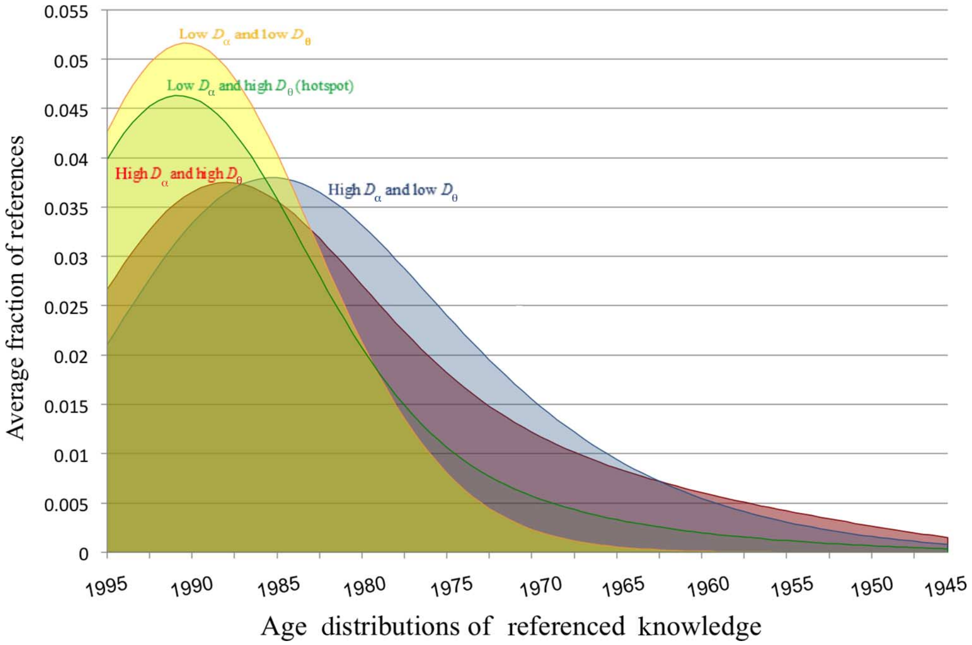

Fig. 2. Distributions of the age of references. The plot shows the characteristic age distributions that correspond to the four quadrants shown in Fig. 1, taking the average distribution for each category among all papers in the WOS published in 1995. The central tendency of the low D_μ and high D_θ, "the knowledge hotspot," distribution includes very recent work with a long, slowly sloping tail into past knowledge. By contrast, the central tendency of the low D_μ and low D_θ distribution is recent work, the central tendency of the high D_μ and high D_θ distribution is relatively old work, and the central tendency of the high D_μ and high D_θ distribution is to cite relatively evenly over past knowledge.

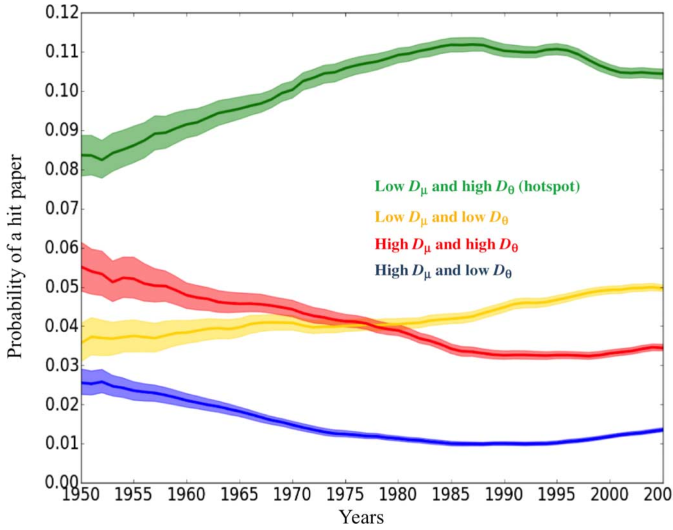

Fig. 3. Increasing dominance of the knowledge hotspot for predicting hit papers in science. Examining scientific papers over time shows that papers referencing work in the "low D_μ and high D_θ" distribution (that is, the knowledge hotspot) have consistently had the highest impact during the past 55 years. The probability of being a hit paper is more than twice the expected background rate, and the gap in citation impact between papers in the hotspot and those outside the hotspot is growing over time. After 1960, only papers that referenced work with certain age distributions, that is, belong to the hotspot, were associated with high-impact research at a rate that exceeded the rate expected by chance.

at the 1st, 10th, 25th, and 50th percentiles of citations, as the log of the number of citations a paper acquires in its first 8 years after publication, citations acquired over a paper's lifetime, a paper's PageRank (*18*), or whether a paper receives the bulk of its citations long after its year of publication, that is, "sleeping beauties" (*17*).

The knowledge hotspot and patenting impact

Like scientific papers, patent impact is measured by citations received, and their references indicate the prior literature the new patent is based on (*3*, *20*). In patenting, the patent examiner's official obligation is to augment authors' citations by citing relevant work authors miss and minimizing irrelevant citations and strategic citations (*21*). Further, patent examiners assign references after seeing the submitted patent. Thus, their retrospective citation process helps identify the knowledge space applicable to a patent.

Patents have the same hotspot-hit relationship as scientific papers. Figure 6 indicates that patents that are in the hotspot are significantly overrepresented at the 5% level of impact. Like papers, patents in the hotspot reference some recent patents (low average D_μ = 6.08, SD = 1.75) and papers of a relatively wide variation of age (high D_θ = 0.98, SD = 0.22). Figure 6 shows that the same hotspot-hit paper relationship holds on an annual basis for patents. Figure 6 indicates that the hotspot is consistently overrepresented in relation to hit patents in 95 to 100% of the patenting subfields over our time frame of 30 years. Tables S11 and S12 present fixed-effects regressions confirming these results net of controls, indicating that the hotspot-hit relationship is robust for inventors and technology and that two critical knowledge creation domains share surprisingly similar and nearly universal patterns relating the age distribution of the referenced literature in a work and a work's probability of being a hit.

Search and the hotspot

The nearly universal benefit linked to the hotspot in science and patenting raises a question as to the factors related to authoring work in versus out of the hotspot. Previous work has found a link between teamwork in science and a paper's citation impact (*2*, *22*, *23*). However, the mechanisms behind the team effect and whether the same scientist performs better working alone or in teams remain unknown (*9*, *12*, *24*–*26*). One

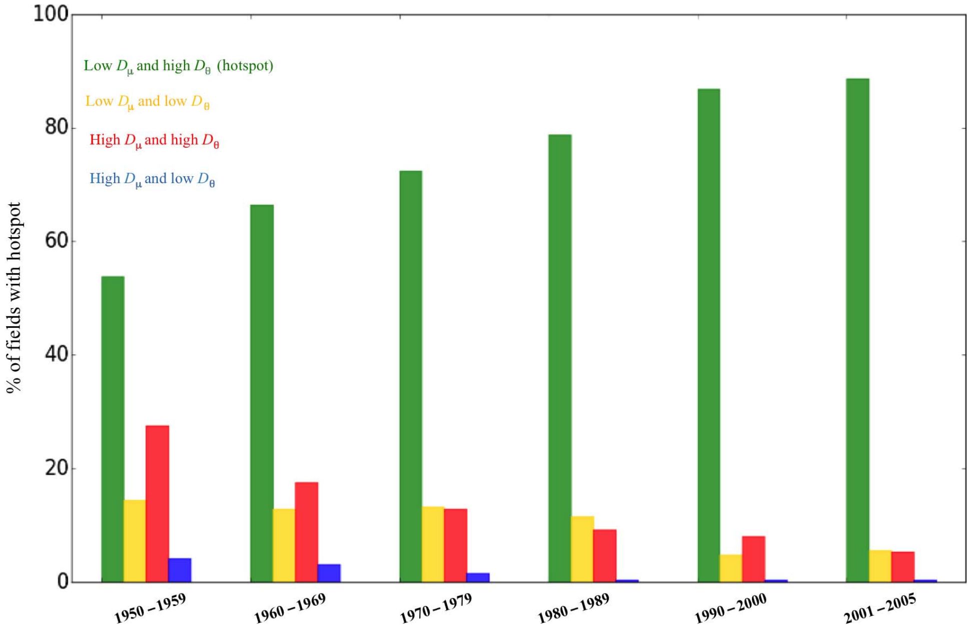

Fig. 4. Knowledge hotspot dominates high-impact science on a field-by-field basis. Disaggregating science into 171 separate science and engineering fields, 54 social science fields, and 27 humanities fields, the histograms indicate the fraction of all fields, where the knowledge hotspot predicts hit papers. In 1990–2000, almost 90% of the 252 fields showed the hotspot-hit link (P < 0.0001, two-tailed binomial test).

Table 1. Probability of being in the top 5% of citations for scientific papers. Logit regression estimates for three time periods indicate that the strong negative predictive relationship between D_μ and H and the strong positive relationship between D_θ and H shown in Figs. 1 and 5 hold across time, fields, paper, and reference characteristics. BIC model fit statistics "very strongly" indicate that models with D_μ and D_θ significantly and substantively fit the data better than control variable models (see Materials and Methods) [(*25*), p. 139]. Variance inflation factor statistics are 1.25 or 1.21, depending on the decade, and indicate no multicollinearity among the independent variables. ***P < 0.0001, **P < 0.001.

	1980–1989		1990–2000		1950–2000	
	β (SE)	**β (SE)**	**β (SE)**	**β (SE)**	**β (SE)**	**β (SE)**
D_μ	−0.195*** (0.0008)	−0.185*** (0.001)	−0.203*** (0.0006)	−0.179*** (0.001)	−0.157*** (0.0004)	−0.179*** (0.0006)
D_θ	1.691*** (0.007)	1.329*** (0.010)	1.559*** (0.0056)	1.410*** (0.007)	1.776*** (0.004)	1.367*** (0.10)
Reference-level controls						
P (Interdisciplinarity)		1.954*** (0.024)		1.892*** (0.021)		1.909*** (0.030)
A (Novelty)		0.185*** (0.006)		0.177*** (0.005)		0.186*** (0.003)
C (Conventionality)		0.239*** (0.002)		0.255*** (0.002)		0.229*** (0.001)
M (Reference quality)		0.001*** (10^{-04})		0.0006*** (10^{-04})		0.001*** (6.5×10^{-06})
Paper fixed effects						
N (#Authors)		Y		Y		Y
Y (Year)		Y		Y		Y
R (#References)		Y		Y		Y
S (Subfield)		Y		Y		Y
Obs.	3,792,038	3,627,624	6,298,005	6,099,788	13,950,691	13,387,366

conjecture is that collaboration potentially reduces knowledge search and awareness problems that solo authors face (*2*, *27*). We examined authors and inventors who created works on their own and in collaboration with others and tested whether the same author is more or less likely to write papers that are in the hotspot when authoring alone versus coauthoring with others. The data used in this analysis comes from two sources: all Fields Medalists in mathematics and patentees. Fields Medalists offer a conservative test of the collaboration conjecture. If collaboration helps

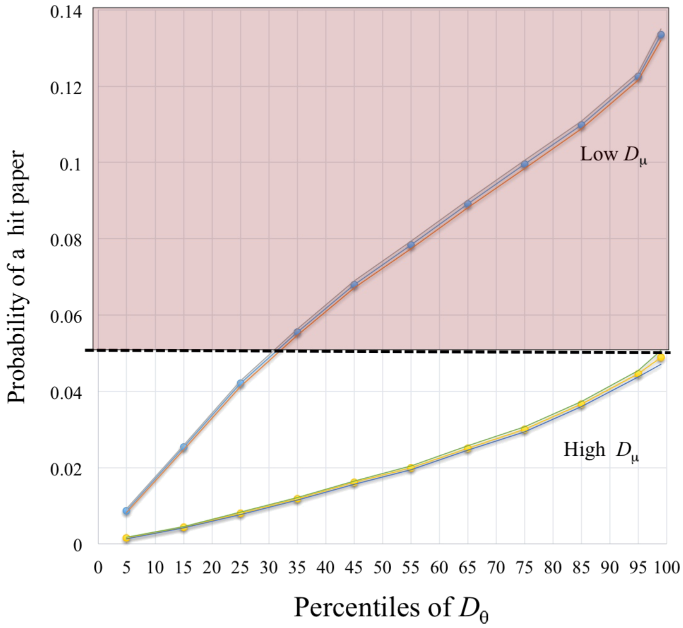

Fig. 5. Probability of a hit paper and combinations of D_μ and D_θ. Estimates are from Table 1 for 1990–2000 with 95% confidence intervals. Combinations of D_μ and D_θ above the dashed line have a probability greater than the 5% background rate expected by chance.

augment the search capabilities of individual scientists, then exceptional scientists may be least likely to receive a boost from collaboration. Patentees provide a test of the collaboration conjecture for all patentees in the U.S. patent database. (Note: We could not analyze all WOS authors because WOS author names lack disambiguation.) Using these data sets, we implement regressions that include fixed effects for each individual author. This approach estimates a within-subject design (rather than a between-subject design), treating each author as her own control case and accounting flexibly for the author's fixed characteristics (for example, IQ, training, and personality). The regression estimates the increase in the probability of a given author producing a paper in the hotspot versus outside the hotspot as a function of whether a given author worked alone or collaborated with others. The regression additionally has fixed effects controls for field, year, and number of references. (See Methods for the regression model and variable construction details.)

Figure 7 graphically presents the results for Fields Medalists. More than 80% of Fields Medalists are significantly more likely to have papers in the hotspot when coauthoring than when writing a paper alone, a relationship unlikely to happen by chance ($P < 0.00009$, binomial test). Consistent with our main effects, Fields Medalists' papers in the hotspot are twice as likely to be their most cited papers on average, reinforcing our general result. Examining patentees, we found a comparable increase in the probability of being in the hotspot associated with collaboration. Collaboration significantly ($P < 0.0001$) improves the likelihood of inventors writing patents that are in the hotspot (table S13).

DISCUSSION

Scientists and inventors prospect an ever-expanding knowledge space in pursuit of new ideas and discoveries. More knowledge suggests more creative material to draw upon, but scientists and inventors are limited in their capacity to search through the knowledge space. This search

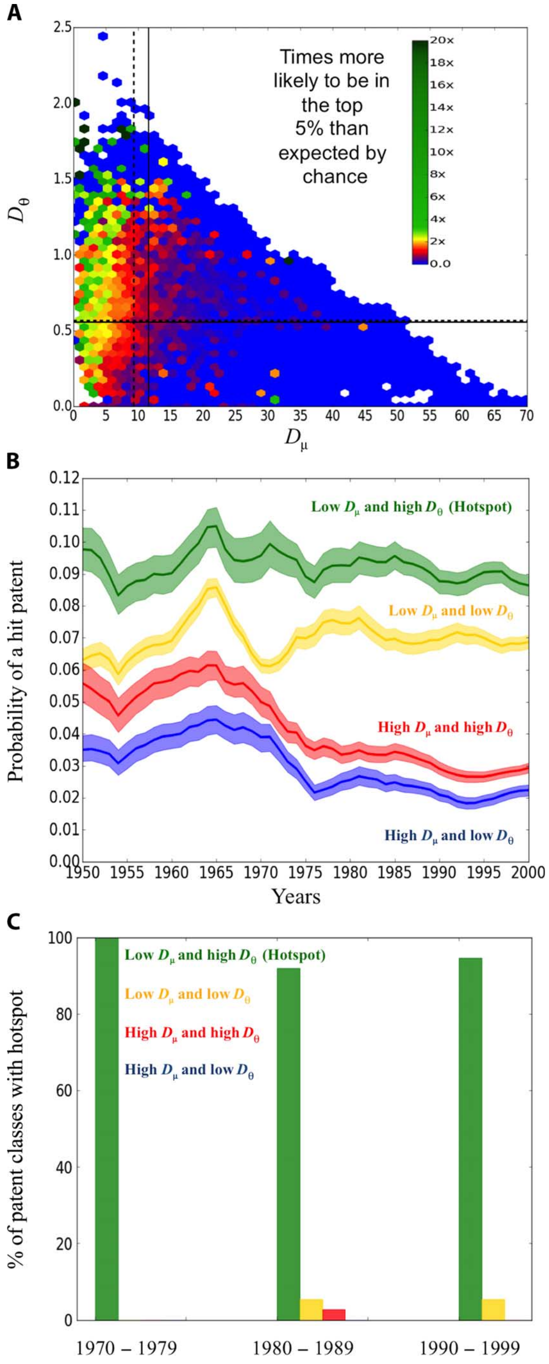

Fig. 6. The dominance of the hotspot for predicting hit patents. (**A**) Knowledge hotspot predicts high-impact technology. Patents that are in the hotspot are more than two times more likely to be hits than the background rate of 5% (data shown are for the year 1995, $N = 103{,}700$ patents). These papers cite prior work that has a low mean age, D_μ, and a high age variance, D_θ, relative to other papers in their field. Notably, 75% of patents are outside the hotspot and display a probability of being a hit that is no greater than expected by chance. Solid lines and dotted lines are population means and medians of D_μ and D_θ. (**B**) Increasing dominance of the knowledge hotspot in patenting. Examining patents on a year-by-year basis shows that patents in the hotspot have consistently had the highest probability of a hit during the past 50 years. (**C**) Knowledge hotspot dominates high-impact patenting on a field-by-field basis. Across 95% patent subfields, patents in the hotspot are more likely to be hits than those based on other ages of information. Between 1990 and 1999, patents in the top 5% of the citation distribution are in the hotspot in more than 95% of subfields ($P < 0.0001$, two-tailed binomial test).

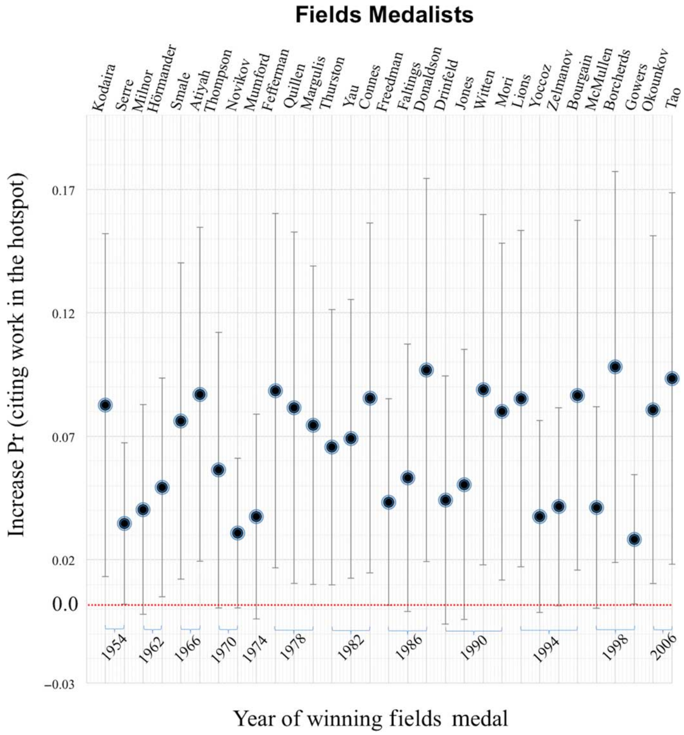

Fig. 7. Collaboration predicts the increased probability of referencing knowledge in the hotspot. Each entry on the *x* axis indicates a different Fields Medalist in mathematics in chronological order of receiving the prize. Values above zero on the *y* axis indicate the difference in the probability of being in the hotspot when a Fields Medalist coauthors versus authors alone. For 26 of 31 Fields Medalists, coauthorship is positively and significantly associated with the authors' chances of being in the hotspot ($P < 0.0009$, binomial test).

trade-off puts a premium on knowing where to search in the literature to discover the most valuable building blocks of new knowledge.

Some theories of knowledge emphasize the importance of using and combining recent ideas in driving breakthroughs, whereas others purport that past knowledge that has withstood the test of time is most valuable (*10*–*12*). Our findings show that each approach is only partly correct. Drawing narrowly on recent ideas does not lead to exceptional impact. Similarly, drawing on vintage knowledge or widely sampled work is associated with an impact no greater than expected by chance.

We find a subtler yet nearly universal pattern that links the age of past knowledge to high-impact work. Our work indicates that a knowledge hotspot characterizes a distribution of the age of prior literature referenced by a paper, relative to the paper's publication year, that is associated with exceptionally high impact in science and technology. Papers and patents in the hotspot reference literature with a low mean age and high age variance relative to a work's publication year. Works in the hotspot more than double their probability of being in the top 5% of impact in their field. Works outside of the hotspot—work centered on recent papers, old papers, or a broad sample of new and old works—do no better than expected by chance. The hotspot's significance is further highlighted by the fact that the highest-impact works across scientific and patenting fields have a similar hotspot-hit relationship. Beyond science and technology, work in progress indicates that the hotspot also reflects the relationship between past and future knowledge in law. In other work, we found that Supreme Court rulings in the United States, Canada, and India that are in the hotspot are overrepresented among the most influential laws (*28*). Last, the hotspot is becoming increasingly predictive of high-impact work over time and now appears in nearly all subfields of science and technology.

The hotspot's generality indicates that there is an age distribution of prior knowledge that is particularly linked to tomorrow's breakthroughs. Future research should begin to investigate what is unique about the knowledge that follows this distribution. Consider two papers written at the same time on the same topic but one paper is in the hotspot and the other is not. What ideas does the former paper have that are not found in the knowledge space searched by the later paper? One conjecture is that the nature of scientific and technological progress involves new knowledge absorbing, replacing, or improving upon prior knowledge. However, these processes may often take time before critical tests can be conducted, debates can be settled, funds can be garnered for addressing the biggest problems and for a large community of scholars to form around the problem. Thus, although a narrow focus on recent literature may offer an opportunity to capitalize on the latest ideas, the research may turn out to be a fad or dead end. At the same time, a narrow focus on vintage work may fail to connect classic ideas with current problems or insights.

Why is it that 75% of the papers and patents are outside the hotspot? We found that one determinant of being in the hotspot is related to collaboration: Authors are more likely to produce work that is in the hotspot when coauthoring than when working alone. Why teams are associated with higher impact work is still an open question and may be related to several explanations that still need to be tested, including a division of labor, collective intelligence, benefits of specialization, positive competition among teammates, myopic search, and social support (*29*). Our findings provide a new explanation for the correlation found between team science and impact with a focus on the role teams play in searching complex knowledge spaces. Amidst these new questions and directions for future work, our findings reveal that the age of information is a remarkably powerful and heretofore unknown predictor of high-impact work in science and technology.

MATERIALS AND METHODS

Data sources used in the analyses

Scientific papers database.

We examined all 28,426,345 research articles indexed in the Thomson Reuters WOS database that were published from 1945 to 2013. The subfield designation of science and engineering (171 subfields), social sciences (54 subfields), and arts and humanities (27 subfields) was defined by the WOS and covers research publications in science and engineering since 1945, social sciences since 1956, and arts and humanities since 1975. These data are described in the Supplementary Materials and are available from Thomson Reuters.

Patents database.

We studied all 5,382,833 patents granted by the USPTO between 1950 and 2010. The data are described in detail in the Supplementary Materials and are available from https://iu.app.box.com/patents/1/779886700/7307669062/1, https://iu.app.box.com/patents/1/779886700/15411270285/1, and https://dataverse.harvard.edu/dataverse/patent. See the Supplementary Materials for further details.

Fields Medalist database.

We collected the data on all 31 mathematicians who won the Fields Medal between 1954 and 2006. This time period allowed us to have at least 8 years of forward and backward citation data for the Fields Medalists' papers. Data are located at http://ams.org/mathscinet/search/author.html?mraurhid=%s&Submit=Search. See the Supplementary Materials for further details.

Measures

High-impact scientific papers and patents ("hits").

We measured a work's H for our main results based on the number of times it was cited in the first 8 years of publication in its respective subfield of science (*6*, *18*, *21*, *22*, *30*). A work with high impact was defined as being in the top 5% of cited works in its specific subfield (that is, 1 of 252 subfields) and year of publication. To test the robustness of our measures, we described many alternative measures and tests of H in the Supplementary Materials, all of which produced results that are in agreement with the ones reported in the main text. The alternative measures were hits defined as being in the top 1, 10, and 25% of the citation distribution, the log of the number of citations a paper accumulates after 8 years of publication and over its lifetime, and a paper's PageRank of citations, that is, papers with a low number of absolute citations but with relatively many cites by hit papers (*18*) and "sleeping beauty" papers (papers that receive the bulk of their citation long after the year of their publication) (*17*). All tests of alternative measures produced confirmatory results.

Age of referenced knowledge.

For each paper and patent, we measured (i) the average age of references, D_μ, and (ii) the COV of the age of references, D_θ. Specifically, D_μ is the mean duration in years between a work's publication year and the publication years of the documents it references (*6*). For example, in fig. S2A, we considered a paper published in 1995 and referencing prior works published in 1990, 1988, 1987, and 1985; this paper has a D_μ of 7.5 (the mean of 5, 7, 8, and 10 years between the publication year and reference years). A work with relatively small D_μ references relatively recent work. In the preceding example, D_θ is 0.308 (SD of 2.16 years normalized by the mean of 7.5 years). A work with a low D_θ references knowledge that is relatively narrowly dispersed around its mean value. Note that a paper's age distribution of references can have a low D_μ and high D_θ whenever the focal paper and most of its references are published close in time, but a few references are published many years before the focal paper. When the focal paper and its references are published close in time, the paper has a low mean age (*31*). When a paper has a low mean age but a small share of its references were published years before the focal paper, the SD and the COV become large, producing papers that have a low mean age and a high age variance, as shown in Fig. 1. About four percent of all papers had a zero difference between their publication year and reference years and were omitted from the analysis. (Please see the Supplementary Materials for a case example, numerical simulations that generalize the case examples, and further details.) For papers published between 1950 and 2000, the mean and SD of D_θ and D_μ are 0.731 and 0.239, and 8.482 and 4.293, respectively. For patents published between 1980 and 2000, the mean and SD of D_θ and D_μ are 0.549 and 0.287, and 12.323 and 8.563, respectively. To test the robustness of measures D_θ, D_μ, and H, we showed the results for alternative measures of the age of information in the Supplementary Materials, all of which are in agreement with the results reported in the main text (tables S1 and S3 to S9).

Methods

Fixed-effects regressions: Predicting citation impact.

For predicting citation impact, the regression for scientific papers takes the form

$$\Pr(H_i) = f\left(D_{i\mu},\ D_{i\theta},\ p_i,\ m_i,\ c_i, a_i, \sum_r \beta_r R_{ri}, \sum_n \beta_n N_{ni}, \sum_f \beta_f S_{fi},\ \sum_y \beta_y Y_{yi}\right)$$

and the regression for patents takes the form

$$\Pr(H_i) = f\left(D_{i\mu},\ D_{i\theta},\ p_i,\ m_i,\ \sum_r \beta_r R_{ri}, \sum_n \beta_n N_{ni}, \sum_f \beta_f S_{fi},\ \sum_y \beta_y Y_{yi}\right)$$

To isolate the effects of our main variables from other predictors of H (*6*, *15*, *19*, *25*, *32–34*), we ran logistic regression models, where we regressed H on our main explanatory variables D_μ and D_θ. Control variables include those specific to a paper or patent (work i) and fixed effects that are specific to categories in the data. The set of control variables varies slightly between papers and patents given data availability, as defined below.

Dependent variable: Citation impact, H ($H = 1$ if a paper or patent is in the top 5% percentile of citations as defined above; 0 otherwise)

Predictor variables: D_μ and D_θ.

Control variables used in the regression analyses.

p_i measures the degree to which work i references prior work from multiple/interdisciplinary fields. To operationalize this variable, we assigned each pair of references in work i's bibliography a value of 1 if the pairs are from the same discipline (as defined by the WOS or USPTO) and 0 otherwise. We then took the average across all reference pairs in a bibliography to compute work i's interdisciplinarity, which varies from 0 to 1. Global means (SDs) are 0.652 (0.208) for papers and 0.655 (0.339) for patents.

a_i measures the degree to which a paper i references prior work that represents novel pairings of prior ideas and is operationalized using the measures described by Evans (*6*). Global means (SDs) are 0.331 (0.470) for papers and not available for patents.

c_i measures the degree to which a paper i references prior work that represents conventional pairings of prior ideas and is operationalized using the measures described by Evans (*6*). Global means (SD) are 4.237 (1.598) for papers and not available for patents.

m_i measures the degree to which work i references highly cited papers (*2*). To control for possible differences in the quality of referenced information, we computed the mean number of citations accumulated by all the references in work i's bibliography. For example, if work i references a total of three papers that have accumulated 10, 5, and 30 citations, then m_i is equal to 15 (45 citations/3 references). Global means (SDs) are 70.561 (144.586) for papers and 8.759 (11.220) for patents.

Fixed-effects controls.

N controls for the number of authors on work i (*18*, *28*). We included indicator variables for one, two, and three or more authors. In the regressions, the omitted indicator variable was for solo authorship.

Y controls for time fixed effects, that is, features of the data that are constant within a year but vary across years such as number of papers published, cohort, annual amount of funding awarded, number of scientists or patentees, and so on. We created an indicator variable for each year, where 2000 was the omitted year for the regression model for papers (1950–2000) and patents (1980–2000).

R controls for the total number of references in work i's bibliography. We created 10 indicator variables (1 = yes; 0 otherwise) for 10 different categories of reference counts. Category 1 has a range of references from 0 to 10, category 2 has a range of references from 11 to 20, etc., with the final category representing 90 references or above. Using 10 equally sized percentile groupings produced the same results. In the regressions, the omitted indicator variable was for category 1. Global means (SDs) are 23.814 (18.439) for papers and 8.656 (10.442) for patents.

S controls for fixed differences across scientific or patenting fields, which include differences between fields in the number of journals, norms of production, topics, and so on. We created an indicator variable for each of the 252 subfields of science and the 36 subfields of patenting. In the regressions, the omitted field indicators were AA (subfield of "acoustics" as indexed in WOS) and 11 (subfield of "agriculture, food, and textiles" for patents).

Fixed-effects regressions: Predicting work in hotspot.

To predict work in the hotspot versus outside of the hotspot, we used the following fixed-effects model

$$\Pr(h_i) = f\left(\alpha n_i + \sum_y \beta_y Y_{yi} + \sum_q \beta_q Q_{qi} + \sum_r \beta_r R_{ri} + \sum_f \beta_f S_{fi}\right)$$

Dependent variable: h_i ($h_i = 1$ if a paper or patent is in hotspot; 0 otherwise)

Predictor variable: n_i ($n_i = 1$ if the paper/patent is solo-authored; 0 otherwise)

Control variables used in the regression analyses.

Q controls for name (in Fields Medalists) or name ID (in patents) fixed effects. We created an indicator variable for every Fields Medalist and patent inventor. This approach means that the regression tells us whether a given individual tends to produce work in the hotspot when that person collaborates with others compared to instances where that same individual works alone. Other control variables in this regression (Y, R, and S) are defined above.

BIC goodness-of-fit statistics for the regression analyses.

BIC statistics were used to interpret the improvement in fit of the regression model when D_μ and D_θ were added to the control variable model (*25*). In all models, the drop in the BIC statistics greatly exceeded 10, indicating that there is "strong evidence" that D_μ and D_θ provide a significantly and substantively better fit to the data than does the control variable model [(*25*), p. 139]. Specifically, the values of BIC goodness-of-fit statistics for analyses with and D_μ and D_θ are as follows:

High-impact papers	
Top 5% of citations in a paper's field	
1950–2000	BIC drops from 4,502,525 to 4,257,045 when D_μ and D_θ are added in the model
1990–2000	BIC drops from 2,001,497 to 1,884,133 when D_μ and D_θ are added in the model
1980–1989	BIC drops from 1,213,883 to 1,147,347 when D_μ and D_θ are added in the model
Top 1% of citations in a paper's field	
1950–2000	BIC drops from 1,427,176 to 1,141,414 when D_μ and D_θ are added in the model
1990–2000	BIC drops from 542,014 to 508,325 when D_μ and D_θ are added in the model
1980–1989	BIC drops from 324,438 to 305,732 when D_μ and D_θ are added in the model
High-impact patents	
Top 5% of citations in a patent's field	
1980–2000	BIC drops from 589,218 to 582,978 when D_μ and D_θ are added in the model
1990–2000	BIC drops from 399,672 to 395,771 when D_μ and D_θ are added in the model
1980–1989	BIC drops from 225,631 to 223,843 when D_μ and D_θ are added in the model
Top 1% of citations in a patent's field	
1980–2000	BIC drops from 160,756 to 159,132.6 when D_μ and D_θ are added in the model
1990–2000	BIC drops from 105,391 to 104,316 when D_μ and D_θ are added in the model
1980–1989	BIC drops from 68,556 to 66,992 when D_μ and D_θ are added in the model

continued on next page

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/4/e1601315/DC1

section S1. Data sets

section S2. D_μ and D_θ distributions

section S3. Alternative measures of D_θ produce equivalent results

section S4. Simple null model of referencing with respect to the age of a publication or patent

section S5. Demonstration case

section S6. Robustness checks for alternative measures of being a hit beyond the top 5% (top 1, 10, 25, and 50%) for papers

section S7. Further robustness checks

section S8. Standardized coefficients of papers

section S9. Regression analyses of patents

section S10. Predicting referencing in the hotspot

section S11. BIC statistics for supplemental regression analyses

fig. S1. Reference age distributions.

fig. S2. Switching references in the null model.

fig. S3. Expected and observed distributions of reference age distributions.

table S1. Alternative measures of D_θ produce consistent results.

table S2. Demonstration case of search and impact.

table S3. Probability of being in the top 1% of citations for papers.

table S4. Probability of being in the top 10% of citations for papers.

table S5. The probability of being in the top 25% of citations for papers.

table S6. The probability of being in the top 50% of citations for papers.

table S7. The probability of being a sleeping beauty paper at different levels of citations.

table S8. Fixed-effects ordinary least squares regression estimating the relationship between D_μ and D_θ and the citations acquired in the first 8 years after publication and for all citations over a paper's lifetime.

table S9. Approximate PageRank analysis.

table S10. Standardized coefficients of the probability of being in the top 5% of papers.

table S11. Probability of being in the top 5% of citations for patents.

table S12. Probability of being in the top 1% of citations for patents.

table S13. Probability of a paper referencing work in the knowledge hotspot for coauthors versus solo authors.

REFERENCES AND NOTES

1. K. Borner, *Atlas of Knowledge* (MIT Press, 2014).
2. B. Uzzi, S. Mukherjee, M. Stringer, B. F. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
3. L. Fleming, Recombinant uncertainty in technological search. *Manage. Sci.* **47**, 117–132 (2001).
4. T. T. Hills, P. M. Todd, D. Lazer, A. Redish, I. D. Couzin; Cognitive Search Research Group, Exploration versus exploitation in space, mind, and society. *Trends Cognit. Sci.* **19**, 46–54 (2015).
5. J. G. Foster, A. Rzhetsky, J. A. Evans, Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
6. J. Evans, Electronic publication and the narrowing of science and scholarship. *Science* **18**, 395–399 (2008).

7. D. Lazer, A. Friedman, The network structure of exploration and exploitation. *Adm. Sci. Q.* **52**, 667–694 (2007).
8. D. Kuksov, J. M. Villas-Boas, When more alternatives lead to less choice. *Mark. Sci.* **29**, 507–524 (2010).
9. B. F. Jones, The burden of knowledge and the death of the Renaissance man: Is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
10. T. S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago, 1962).
11. R. K. Merton, Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proc. Am. Philos. Soc.* **105**, 470–486 (1961).
12. D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
13. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
14. J. M. Kleinberg, Authoritative sources in a hyperlinked environment. *ACM* **46**, 604–632 (1999).
15. M. J. Stringer, M. Sales-Pardo, L. A. N. Amaral, Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *J. Am. Soc. Inf. Sci. Technol.* **61**, 1377–1385 (2010).
16. S. Wuchty, B. F. Jones, B. F. Uzzi, The increasing dominance of teams in the production of knowledge. *Science* **316**, 1036–1039 (2007).
17. Q. Ke, E. Ferrara, F. Radicchi, A. Falmmini, Defining and identifying Sleeping Beauties in science. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7426–7431 (2014).
18. P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with Google's PageRank algorithm. *J. Informetr.* **1**, 8–15 (2007).
19. C. J. Lortie, L. Aarssen, J. N. Parker, S. Allesina, Good news for the people who love bad news: An analysis of the funding of top 1% most highly cited ecologists. *Oikos* **121**, 1005–1008 (2012).
20. B. H. Hall, A. B. Jaffe, M. Trajtenberg, Market value and patent citations. *Rand J. Econ.* **36**, 16–38 (2005).
21. R. Lampe, Strategic citations. *Rev. Econ. Stat.* **94**, 320–333 (2012).
22. R. Guiméra, B. Uzzi, J. Spiro, L. A. N. Amaral, Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
23. K. Börner, N. Contractor, H. Falk-Krzesinski, S. Fiore, K. Hall, J. Keyton, B. Spring, D. Stokols, W. Trochim, B. Uzzi, A multi-level systems perspective for the science of team science. *Sci. Transl. Med.* **2**, 49cm24 (2010).
24. B. F. Jones, E. J. Reedy, B. A. Weinberg, Age and scientific genius, in *The Wiley Handbook of Genius*, D. K. Simonton Ed. (Wiley-Blackwell, 2014).
25. A. E. Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
26. S. F. Lu, G. Z. Jin, B. Uzzi, B. F. Jones, The retraction penalty: Evidence from the Web of Science. *Sci. Rep.* **3**, 3146 (2013).
27. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
28. R. Whalen, S. Mukherjee, B. Uzzi, Common law evolution and judicial impact in the Age of Information. *Elon Law Rev.* (2017).
29. Committee on the Science of Team Science, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, *Enhancing the Effectiveness of Team Science*, N. J. Cooke, M. L. Hilton, Eds. (National Academy of Science, 2015).
30. B. F. Jones, S. Wuchty, B Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
31. E. Matricciani, The probability distribution of the age of references in engineering papers. *IEEE Trans. Prof. Commun.* **34**, 7–12 (1991).
32. P. Pirolli, Rational analyses of information foraging on the web. *Cognit. Sci.* **29**, 343–373 (2005).
33. P. Azoulay, J. L. Furman, J. L. Krieger, F. E. Murray, Retractions. *Rev. Econ. Stat.* **97**, 1118–1136 (2015).
34. M. E. J. Newman, Prediction of highly cited papers. *Europhys. Lett.* **105**, 28002 (2014).

Acknowledgments: We thank A. Pah, S. Saavedra, C. Huepe, A. Horvát, and J. Poncela for their helpful comments. **Funding:** This research was funded by grants from the Northwestern University Clinical and Translational Sciences Institute, the Northwestern University Institute for Complex Systems, and the Multidisciplinary University Research Initiative–Defense Advanced Research Projects Agency (DARPA) (grant W911NF-15-1-0577). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the social policies, either expressed or implied, of DARPA or the U.S. government. **Author contributions:** S.M. conducted data collection and analysis, designed analyses, and wrote the paper. D.M.R. conducted data analysis, designed analyses, and wrote the paper. B.J. designed analyses and wrote the paper. B.U. conducted data collection and analysis, designed analyses, wrote the paper, and received grant funding. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials and are available online through the WOS. Additional data related to this paper may be requested from the authors.

Submitted 10 June 2016
Accepted 23 February 2017
Published 19 April 2017
10.1126/sciadv.1601315

Citation: S. Mukherjee, D. M. Romero, B. Jones, B. Uzzi, The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Sci. Adv.* **3**, e1601315 (2017).

Science Advances

The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot

Satyam Mukherjee, Daniel M. Romero, Ben Jones and Brian Uzzi

Sci Adv **3** (4), e1601315.
DOI: 10.1126/sciadv.1601315