Impact of geographical distance on acquiring know-how through scientific collaboration

Frank van der Wouden[a,b,c] **and Hyejin Youn**[b,c]

[a]Department of Geography, University of Hong Kong, Pokfulam Road, Hong Kong; [b]Management & Organizations, Kellogg School of Management, Northwestern University, 2211 Campus Dr, Evanston, Illinois 60208, United States of America; [c]Northwestern Institute on Complex Systems, 600 Foster Street, Evanston, Illinois 60208, United States of America

This manuscript was compiled on February 26, 2020

Individuals who collaborate locally are more likely to learn from one another than those collaborating non-locally. We examine the publication records and knowledge portfolios of almost 1.7 million scholars who published a co-authored paper and subsequently published a single-authored paper. We investigate to what extent the geographical distance between the collaborators impacts the probability of them learning through collaboration. We find evidence that local collaboration is associated with a learning premium of between 40% and 85%. Surprisingly, this learning premium increased over time, despite advances in communication and transportation technologies that supposedly abolish the friction of geographical distance. In addition, we find that the learning premium from local collaboration is greater in individuals in their early career stage than in mid or late career stages. Furthermore, individuals from lower ranked institutions are most likely to learn from collaborations. Our findings also indicate that increasing geographical distance between collaborators hits the collaborative learning of those in STEM-type fields the hardest. We observe highly significant positive effects from local collaboration on learning even after controlling for confounding factors and using matched data in the statistical models. These results suggest that geography plays a key role in mitigating the impact of learning through collaboration. This has important implications on innovation policy, the structure of effective research teams, and the broader processes of knowledge production and diffusion.

Collaboration | Know-how | Learning | Distance | Publications

Humankind's ability to survive across diverse environments on the face of the Earth is largely attributed to our ability to learn from other humans through observation and imitation (1). This social learning is what enables humans to collaborate and to share information, skills, and practices that increase our ability to adapt to local conditions and changing environments in places that we inhabit. At the population level, this ability allows the accumulation of information, knowledge, and know-how across generations. Instead of continually re-inventing the wheel, wasting time and resources, humans can focus their attention on tackling new challenges (2).

Social learning is indeed a key mechanism through which know-how can diffuse across human populations in both time and space. In modern society, know-how refers to our ability to perform tasks 'smoothly and efficiently' (3). In this sense, know-how differs from conventional knowledge. Whereas knowledge can be codified as 'building-blocks' of the 'what', know-how refers to the practices and routines required to successfully utilize these building blocks to perform any task. For example, having knowledge of certain aspects of riding a bicycle, such as gravity or rolling resistance, is not sufficient for one to ride a bicycle. Rather, cycling requires the accumulation of practice and the familiarisation of an action for it to become routine. These practices and routines are tacit in nature, embodied in individuals and organizations, and hence embedded in physical spaces. To diffuse the practices and routines that constitute the know-how of cycling, social learning processes are imperative (4). Repeated face-to-face interactions with our siblings and parents riding bicycles enables us to observe and imitate how to ride a bicycle successfully. Thus, the diffusion of know-how through social learning has a distinct geographical element.

It can be argued that recent advances in information, telecommunication, and transportation technologies free us from the friction of geographical distance in socio-economic activities, and relax the coupling of acquiring know-how through social learning and geographical proximity, resulting in the famous 'death of distance' (5) and 'flat world' (6). Obvious examples include telephone, email, and Internet, which facilitate efficient, affordable, and near-instant communication across large distances. Global transportation has significantly reduced travel times, as an increasing number of cities are now connected by direct flights. Information and telecommunication investments in fiber optics and advanced video technologies in conference rooms allow real-time simultaneous digital face-to-face interactions with other people in different locations. When it comes to knowledge diffusion, distance is death.

At the same time, scholars point out the omnipresence of spatial tension in social and economic structures. Global trends of increasing business travel demonstrate the lasting importance of real face-to-face interactions in business (7–9). At global, national, and regional scales, economic activities are highly concentrated in space, giving rise to technology

Significance Statement

Collaborating locally is more likely to result in learning compared to collaborating non-locally. This is important because people are increasingly collaborating within firms, academia and governments. We showed that collaboration is a great platform that allows know-how to be transmitted, and that this transmission is enhanced by geographical proximity. Despite increasing technological capacity to communicate non-locally, this local learning premium has actually increased over recent years. These findings are important for the structure of research teams and the design of policies aimed at diffusing know-how.

FVDW and HY designed the research, analyzed the data and wrote the manuscript.

The authors declare no conflicts of interest.

[2]Correspondence should be addressed to Frank van der Wouden. E-mail: fvdw@hku.hk

districts and regional pockets of economic specialization (10–12). High rents ought to be a disincentive to agglomeration, yet the rapid growth of large cities suggests that distance is far from dead. The benefits of co-location have long been discussed in economic geography, and they ring especially true for innovation (13). The benefits from geographical proximity appears to be unaffected by technological advance (14–16). Distance matters for knowledge diffusion.

Do humans still require geographical proximity to acquire know-how through interaction, observation, and imitation of others? Advocates of the 'death of distance' argue that acquiring know-how through social learning processes can be decoupled from geographical co-location. Others argue that distance still matters because transmission of know-how requires repeated face-to-face interactions and trustful relationships—ingredients only facilitated by geographical co-location, regardless of advancements in information, communication, and transportation technologies (17–20). Although acquiring know-how is central to many theories on human evolution (2), firm competitiveness (3, 19), and economic development (21, 22), the conundrum of geographical distance on the acquisition of know-how through social learning in the age of advanced technologies remains largely unanswered.

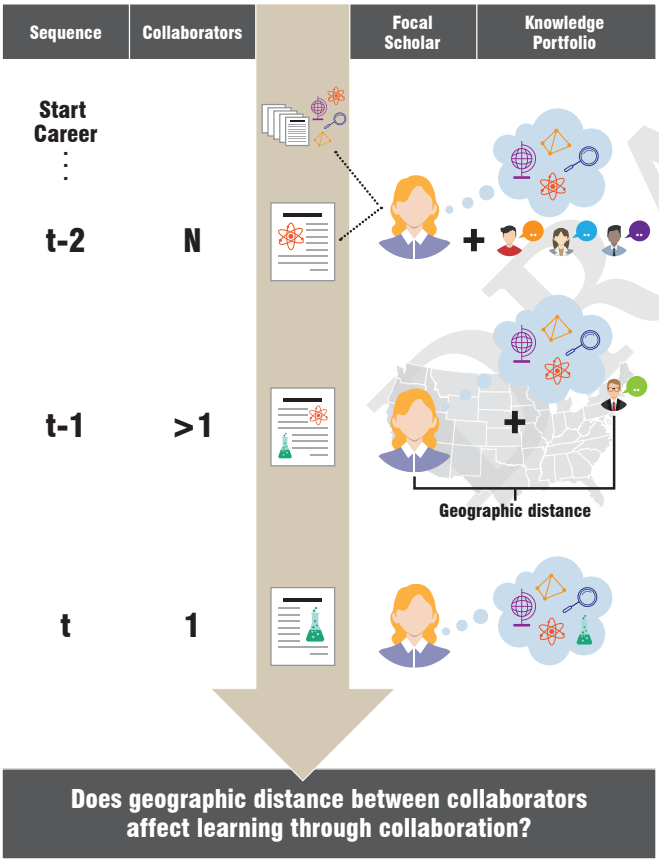


Fig. 1. Sampling procedures and measuring acquisition of know-how (green flask) during collaboration.

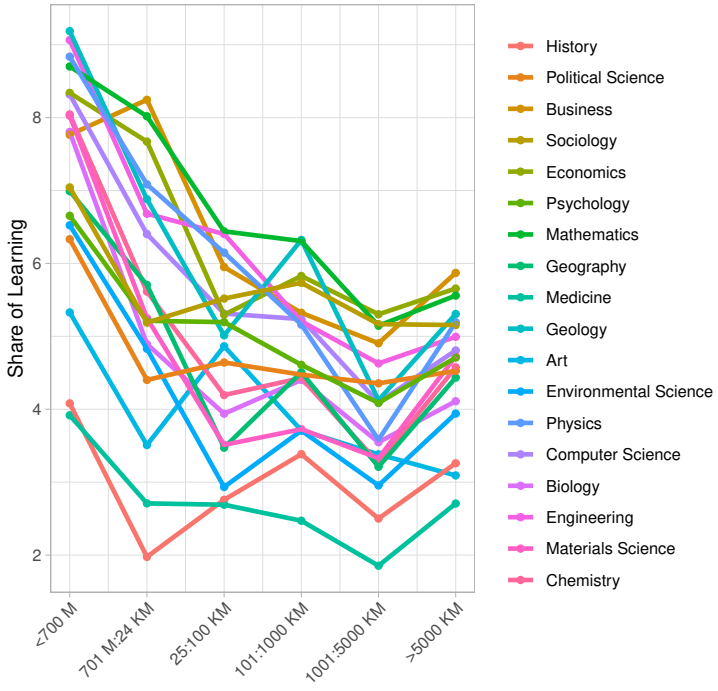
To answer this question, we examine the production of scientific knowledge and whether collaborations represent an important opportunity for social learning and the acquisition of know-how. More precisely, we track scholars over time and examine the impact of geographical distance between collaborators on their ability to acquire know-how. Our data comes from the Microsoft Academic Graph (MAG) (23) , which holds information on over 62 million academic papers linked to more than 200 million disambiguated (co-)authors. Microsoft Research uses natural language processing algorithms to classify each paper into 'Fields of Study' (FOS) based on their topics, fields, and methods. By following a scholar's career over time, a cumulative knowledge portfolio can be constructed that records the FOS associated with each scholar.

From this database, we identify 'focal scholars' who have published a sequence of three consecutive papers that satisfy certain criterea. The first paper in the sequence is used to construct the pre-collaboration knowledge portfolio of the scholar. The second in the sequence is a collaborative paper. The third in the sequence is a single-authored paper. We can extrapolate that the scholar has acquired know-how through a collaboration when the single-authored paper contains at least one FOS occurring in the collaborative paper, but not in the focal scholar's pre-collaboration knowledge portfolio. We presume that this know-how is acquired during the process of collaboration with the co-author(s) and/or directly from the co-author(s) (Figure 1). Note that the third paper needs to be single-authored so that we can examine whether know-how is acquired by the focal scholar and not potentially contributed by a co-author.

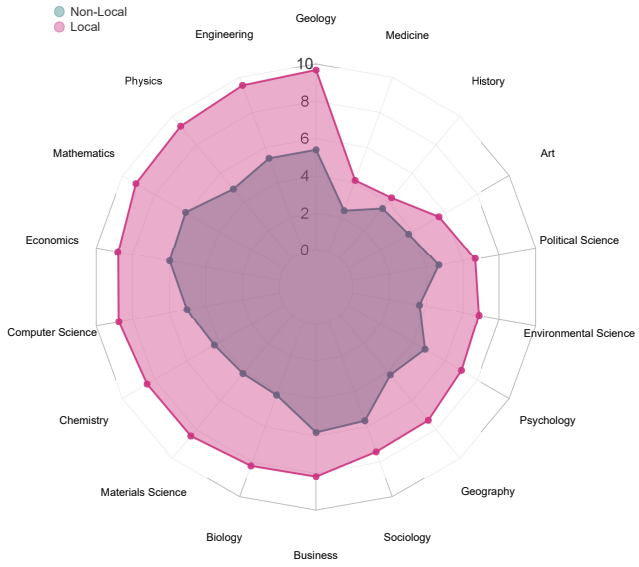
Obviously, scholars constantly learn and generate new know-how and interests not recorded in the (collaborative) academic publishing system. How, why, and when scholars switch or gain new research interests is another ongoing and fascinating research topic (24–28). Here, we strictly focus on the impact of geographical distance on acquiring know-how by looking at the collaborative scholarship of academic papers. This is important, because recent evidence demonstrates that knowledge production is becoming increasingly complex, driving human specialization and collaboration (29–31). Understanding how collaboration can be best organized benefits knowledge production and society at large.

An important issue in our research is to what extent can one assign the 'acquiring of know-how' as an effect of a collaborative co-authorship rather than scholars' natural tendency to explore new interests and ideas. For example, consider a focal scholar starting to explore a new field F with another co-author also inexperienced in field F. They publish a joint paper in this field. If the focal scholar subsequently publishes a single-authored paper on field F, it is difficult to assign whether learning effects are due to the previous collaboration. The focal scholar could have published a similar paper in field F without the co-author. Although this could be true, we assume that during every collaborative co-authorship, the authors acquire know-how from each other. In this research paper, we examine whether the acquisition of know-how is greater when co-authors are in close geographical space.

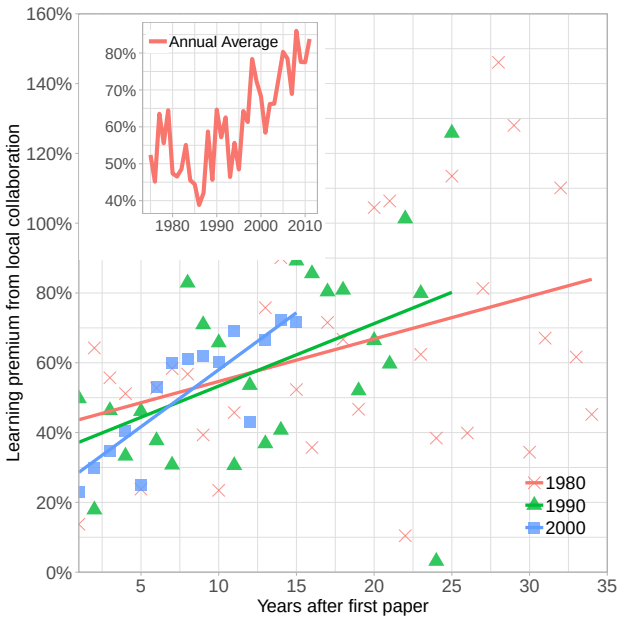
We investigate the acquisition of know-how through collaborations among scholars along several dimensions. Considering globalization and technological advances, our main focus is to examine whether geographical distance between collaborators affects the likelihood of acquiring know-how and how this changes over time. If technological advances can substitute for geographical proximity, then the impact of distance between collaborators should not affect know-how acquisition and any impacts should diminish over time. We then examine whether



(a)



(b)



(c)

Dependent variable: Learning through Collaboration				
	Estimate	Std. Error	t value	p-value
(Intercept)	-3.65	0.33	-11.03	0.00
Local Collaboration	0.30	0.01	42.81	0.00
Number of Coauthors	-0.22	0.00	-79.81	0.00
Number of FOS	0.16	0.00	125.10	0.00
Affiliation Rank 2	0.14	0.01	12.69	0.00
Affiliation Rank 3	0.22	0.01	20.36	0.00
Affiliation Rank 4	0.41	0.01	38.74	0.00
Time Difference	0.02	0.00	36.43	0.00
Career Stage Mid	-0.76	0.01	-78.71	0.00
Career Stage Late	-1.20	0.01	-82.90	0.00

(d) Average treatment effect of Local Collaboration on probability of learning through collaboration. For Rank and Career State the reference category is Affiliation Rank 1 and Career Stage Early. Year and discipline fixed effects are included in model and reported in SI

Fig. 2. The probability of learning through collaboration drops with geographical distance between collaborators and is observed across all academic disciplines (a). Local collaboration is associated with greater learning probabilities across all disciplines (b). There is a learning premium from collaborating locally (c), regardless of cohort and it has increased with time (inset c). The benefits from local collaboration on learning remain positive and statistically significant when controlling for other factors and using matched data (d).

the acquisition of know-how differs across disciplines, team sizes, academic ranks, career stages and time since collaboration. We use coarsened exact matching techniques (32–34) in combination with logistic regression models to statistically estimate the average treatment effects of distance on acquiring know-how. In addition, we train a supervised machine learning model to examine whether distance impacts learning when controlling for complex interactions and non-linear relationships between predictors. We then analyse our results by linking our findings to current debates. The evidence presented from the descriptive analyses, statistical models, and machine learning model suggests that geographical distance negatively impacts know-how acquisition, regardless of technological advances.

Results

Acquiring know-how from collaborations occurred only in a small number of cases in our sample. Figure 3 shows that the annual counts of individuals who acquired know-how through collaboration increased between 1975 and 2015. This was mainly driven by the increasing size of academic publishing over the same period, with more papers produced in the 2000s than ever before. Meanwhile, the annual count of sampled scholars who did not acquire know-how through collaborations increased at a faster rate. As a consequence, the share of individuals who acquired know-how through collaborations dropped from roughly 10% in the 1970s to below 5% after 2010. These findings suggest that acquiring know-how through collaborations is becoming increasingly rare over time.

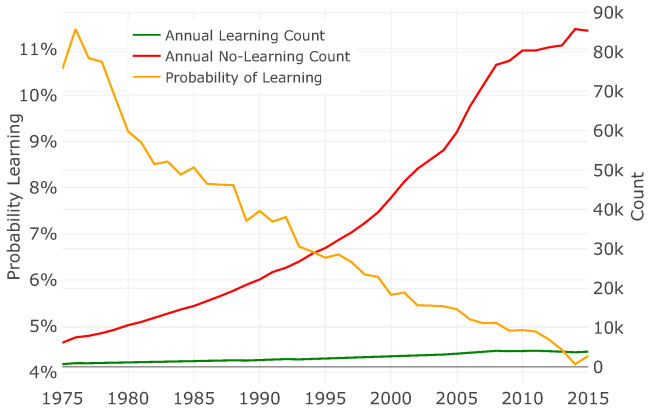


Fig. 3. Count and probability of learning over time.

The average geographical distance between co-authors of academic papers nearly doubled between 1975 and 2015. The average distance between co-authors was 1,000 km in 1975, linearly increasing to almost 2,000 km in 2015. Part of this increase was associated with a decrease in the likelihood of co-authorship with local scholars (Figure 4). We consider a collaboration to be local when the average geographical distance between co-authors is equal or less than 700 m—roughly a 10-minute walk (see Materials and Methods). The percentage of local co-authorships decreased from 75% to less than 60% between 1975 and 2015. Although the majority of co-authorships were still local in 2015, the trend of the past 40 years suggests this will soon not be the case. Part of the reason may be due to the trend of increasing team sizes in academic publishing. The likelihood that a suitable team-member is sourced from a rather finite pool of local scholars decreases with team size. Another possible explanation is that, indeed, advances in telecommunication and transportation technologies have made it easier to communicate with non-local scholars and engage in collaborative relationships.

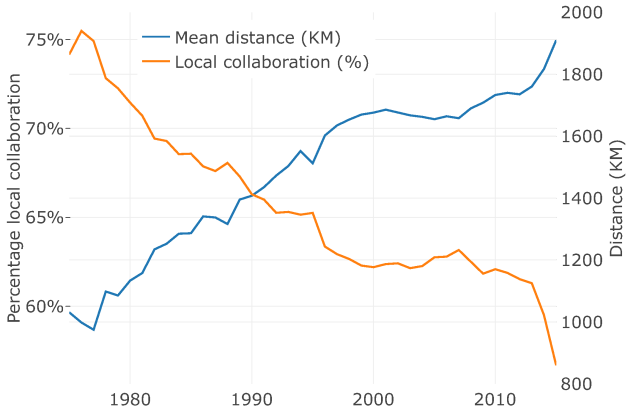


Fig. 4. Probability of learning by local and non-local collaboration.

Does geographical distance impact the probability of acquiring know-how during a collaboration? The key findings of this paper are presented in Figure 2. The geographical distance was found to negatively affect the probability of learning across academic disciplines (Figure 2A). The geographical distance between collaborators had the strongest negative impact on learning in Chemistry, Materials Science, and Engineering disciplines (see Figure 9 in SI), whereas distance had the smallest negative impact in History, Political Science, and Business. A possible explanation is that natural sciences and STEM-type fields rely more on (geographically localized) instruments and equipment than most fields in social sciences and humanities.

The share of learning was greater in local rather than non-local collaborations across all academic disciplines, but the differences fluctuated across disciplines. Figure 2B shows that scholars in the field of Geology had the largest share of learning after local collaboration (almost 10%), whereas scholars in Medicine had the lowest share at 3.9%. The greatest absolute difference between the share of learning from local and non-local collaboration was in Physics, followed by Materials Science and Geology, with larger local learning rates of 4.17%, 4.12%, and 4.06%, respectively. The smallest differences in the learning rates between local and non-local collaborations were in History (0.71%), Philosophy (1.04%), and Medicine (1.63%), which also had relatively small learning rates. The differences in the share of learning between local and non-local collaboration were significantly higher for all disciplines ($2.71 < \text{Z-score} < 1297.81$ and P-values < 0.05).

Figure 2C shows the learning premium from local collaborations for the cohorts 1980, 1990 and 2000 over time. The learning premium is the relative percentage difference between the local and non-local annual learning rates. There was a learning premium from local collaboration for all cohorts and throughout the careers. The flattening slope for earlier cohorts suggests that this premium decreased with the stage of career,

as the slope of each cohort was qualitatively similar in first 15 years of a career (see Figure 10 in SI). However, these results should be treated with caution as observations on the 20 years after the first paper were increasingly scarce. The inset in Figure 2C pools all the observations by year and showed that the learning premium for local collaborations increased over time from roughly 50% in the 1980s to around 80% in 2010. Figure 2D shows the results of a multivariate logistic regression model using matched data, which confirmed these findings. After controlling for several co-variates, scholars who engaged in local collaborations were almost 57% (odds ratio of 1.349) more likely to have learned through collaboration than those who engaged in non-local collaborations. Geographical proximity matters for acquiring know-how through collaboration.

The relationship between learning through collaboration and geographical distance differs across discipline, team-size and rank of institution. Figure 5 plots the learning probability across these variables by geographical distance, representing scales for local, metropolitan, regional, within (nation-)state, within continent and intercontinental collaboration. This figure shows that across all disciplines the probability of know-how acquisition is greatest locally and decays almost linearly with distance. Larger team-size is also associated with smaller probabilities of learning. Larger teams might rather be a collection of specialized authors each contributing their expertise instead of a group of scholars intensely interacting. Given resource restrictions of individuals and academia in general, it seems plausible that the interaction density among collaborators declines with increasing team size. In addition, scholars affiliated with lower-ranked institutions have greater probabilities of learning than those affiliated with higher ranked institutions. For scholars in all ranks, probabilities of learning are highest when collaborating locally. Interestingly, inter-continental collaboration ($>$ 5001 KM) has higher probabilities of learning than within continent but outside (nation-)state collaboration ($>$ 1001 : $<$ 5000 KM). This might be explained by the strong geographical localization of new, specialized, complex types of know-how (35). Diffusion of these types of know-how might flow more readily between related specialized locations regardless of their distance, rather than local locations because a certain absorptive capacity (36) or cognitive proximity (37, 38) is required.

The relationship between learning through collaboration and geographical distance differed across disciplines, team sizes, and rank of institution. Figure 5 shows the learning probability across these variables by geographical distance with representative scales for local, metropolitan, regional, within (nation-)state, within continent, and intercontinental collaborations. Across all disciplines, the probability of know-how acquisition was greatest locally, which decayed almost linearly with distance. Larger team sizes were associated with smaller probabilities of learning. Larger teams might be a collection of specialized authors each contributing their own expertise rather than a group of scholars intensely interacting. Given the resource restrictions of individuals and academia in general, it seems plausible that the interaction density among collaborators declines with increasing team size. In addition, scholars affiliated with lower-ranked institutions had higher probability of learning than those affiliated with higher-ranked institutions. For scholars in all ranks, the probability of learning was the highest with local collaborations. Interestingly, inter-continental collaborations ($>$ 5001 km) had higher likelihood of learning than within continent, but was outside of (nation-)state collaborations ($>$ 1001 and $<$ 5000 km). This might be explained by the strong geographical localization of new, specialized, and complex types of know-how (35). Diffusion of these types of know-how might flow more readily between related specialized locations regardless of their distance, because a certain degree of absorptive capacity (36) or cognitive proximity (37, 38) is required.

Scholars collaborating locally had a greater probability of learning than those collaborating non-locally, regardless of the time between the collaborative and single-authored papers. Figure 6 shows the probability of learning in both groups followed a reversed U-shape. The probabilities increased until roughly 42 months after the collaboration. After 42 months, the probability of learning (and sample size) dropped in both groups. This suggests that the single-authored papers from both groups go through similar publication cycles. Geographical distance between collaborators did not seem to impact the publication speed of the single-authored paper.

The probability of acquiring know-how was greatest when scholars were in their early career stage. The career stage was determined as the difference between the year of the first published paper and the single-authored paper. Figure 7 shows that scholars in late career stages tended to have the lowest probability of acquiring know-how from a collaboration. This makes sense because compared to more senior colleagues whose portfolio would be filled with more FOS, the portfolios of scholars in the early career stages are relative empty, thus they would be more likely to collaborate on a project with a FOS not in their current portfolio. Interestingly, the premium for collaborating locally was present across all career stages. The absolute difference between local and non-local collaboration was greatest for early-career scholars, whereas the largest relative increase was for late-career scholars.

Does geographical distance still impact the probability of acquiring know-how through local collaboration after controlling for these co-variates? We use coarsened exact matching techniques (32–34) to match scholars who collaborated locally with similar scholars who collaborated non-locally by team size, affiliation ranking, discipline, year of collaboration, year of single-authored paper, career stage, and time difference between the collaborative and single-authored papers. The outcome of this exercise is a dataset of local and non-local collaborating scholars that looked similar in terms of co-variates. This allowed us to more precisely estimate the effects of distance (local or non-local) on the outcome of acquiring know-how.

Table 2d presents the results from a multivariate logistic regression model using the matched data. Distance negatively impacts the acquisition of know-how, even when controlling for the co-variates. Scholars collaborating locally are almost 57% (odds of 1.35) more likely to learn through collaboration than those collaborating non-locally. With an increase of one additional co-author, the odds of learning through collaboration decreases by 0.80. The number of FOS classifications on the paper is positively associated with learning. This is to be expected because our measure of acquiring know-how relies on whether FOS from the collaborated paper are successfully produced by the focal scholar on the single-authored paper. An additional FOS on the collaborated paper results in an

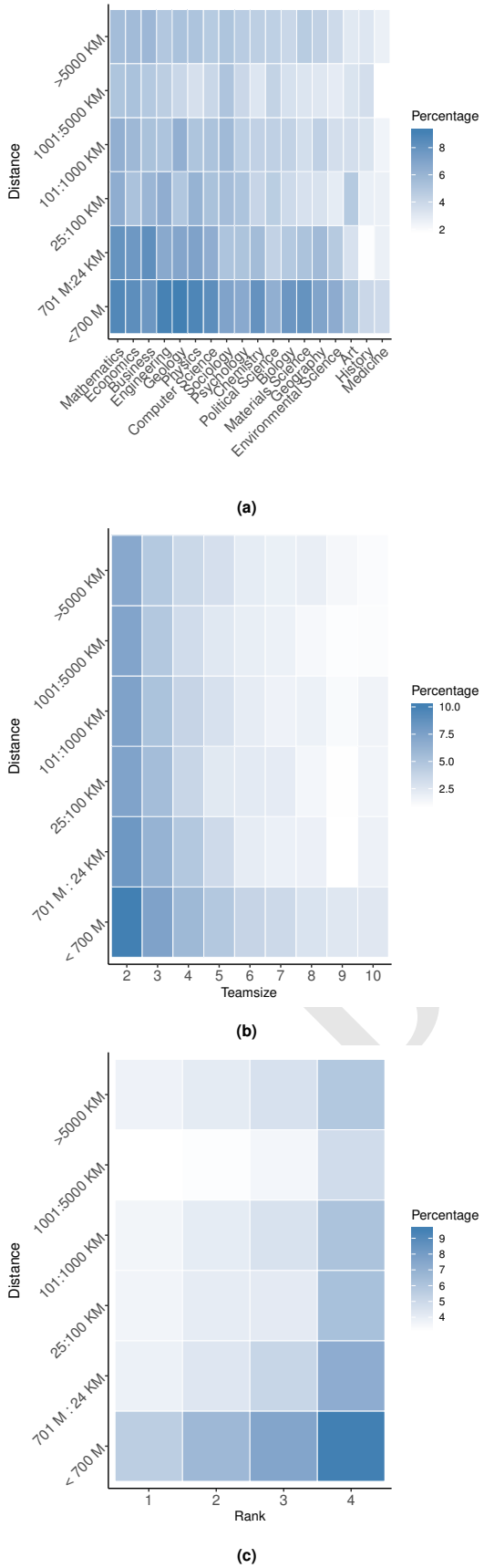


Fig. 5. Percentage of Learning through Collaboration by distance and (a) discipline, (b) team-size and (c) affiliation ranking

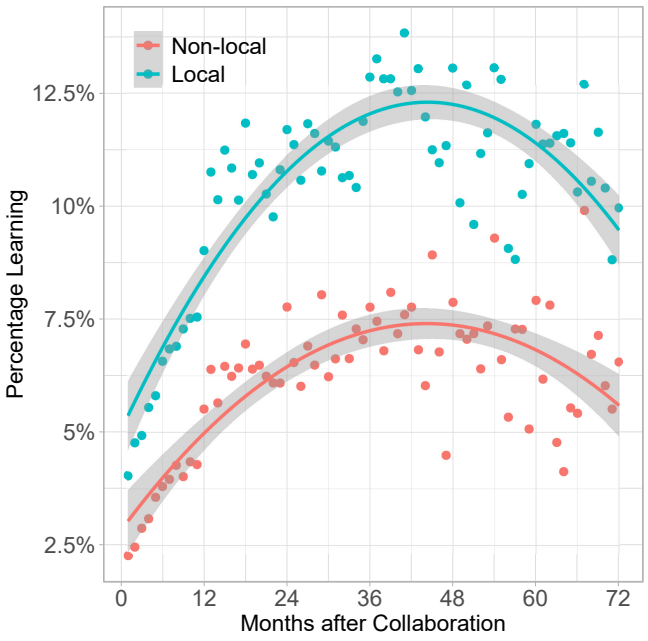


Fig. 6. Learning rate by time since collaboration for local and non-local collaboration.

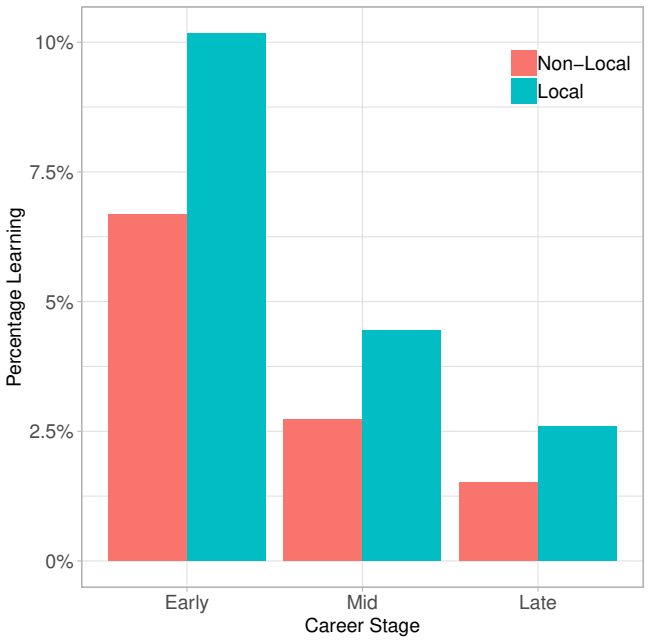


Fig. 7. Probability of learning by career-stage (Early < 11; Mid > 10 & < 21; Late > 20 year).

1.17 increase in the odds of know-how acquisition. Scholars affiliated with lower ranked institutions are more likely to acquire know-how through collaboration than scholars associated with top-ranked institutions. The difference in likelihood of know-how acquisition is greatest between the top and bottom quartile with a difference in odds of 1.50. An increase in time-difference in months between the author's collaborated paper and single-authored paper is also positively associated with acquiring know-how. Finally, scholars in early career-stage are most likely to acquire know-how through collaboration than those in later career stages. Being in the late career stage, versus in early career stage, changes the odds of acquiring know-how through collaboration by 0.29. Affiliation, collaboration and single-authored paper year are included in the model as fixed effects but not shown here. These results support our findings discussed earlier and confirm that scholars who collaborate with local co-authors are more likely to acquire know-how through collaboration than similar scholars who collaborate non-locally. This suggests there is a learning-premium for those who collaborate locally because geographical distance negatively impacts acquiring know-how through collaboration. Results from supervised machine learning models capturing complex interactions and non-linear relationships confirm these findings (see SI).

Figure 2d shows the results from the multivariate logistic regression model using the matched data. Distance negatively impacted the acquisition of know-how, even when controlling for the co-variates. Scholars collaborating locally were almost 57% (odds of 1.35) more likely to learn through collaboration than those collaborating non-locally. The addition of an additional co-author decreased the odds of learning through a collaboration by 0.80. The number of FOS classifications in the paper was positively associated with learning. This is to be expected because our measure of acquiring know-how relies on whether FOS from the collaborative paper was successfully produced in the single-authored paper by the focal scholar. An additional FOS in the collaborative paper resulted in a 1.17 increase in the odds of acquiring know-how. Scholars affiliated with lower ranked institutions were more likely to acquire know-how through collaboration than scholars affiliated with top-ranked institutions. The difference in the likelihood of know-how acquisition was greatest between the top and bottom quartile, with a difference in the odds of 1.50. An increase in time-difference in months between the author's collaborative paper and single-authored paper was positively associated with acquiring know-how. Finally, scholars in their early career stage were most likely to acquire know-how through collaboration than those in later career stages. Being in the late career stage versus early career stage changed the odds of acquiring know-how through collaboration by 0.29. Affiliation, collaboration, and single-authored paper year were included in the model as fixed effects, but are not shown here. These results support our findings discussed earlier and confirm that scholars who collaborate with local co-authors are more likely to acquire know-how through collaboration than similar scholars who collaborate non-locally. This suggests there is a learning-premium for those who collaborate locally as geographical distance negatively impacted acquiring know-how through collaboration. Results from the supervised machine learning models that captured complex interactions and non-linear relationships also confirmed these findings (see SI).

Conclusion

In this paper we examined whether humans still require geographical proximity to acquire know-how through interacting, observing, and imitating one another. Technological advancements in telecommunication and transportation might relax the need for collaborations to require physical co-location. We investigated the probability of acquiring know-how among scholars who engaged in collaborations either locally or non-locally between 1975 and 2015. The evidence presented in this paper indicates that geographical distance negatively impacts know-how acquisition, regardless of technological advances such as the Internet, smartphones, and teleconferencing. Scholars who engaged in local collaboration had higher rates of know-how acquisition compared with those who collaborated non-locally. This learning premium from local collaborations was found to increase with time.

Our findings have important implications for academia. Even though the rate of local collaboration on academic papers is decreasing and the average geographical distance between collaborators is increasing over time, the probability of acquiring know-how through collaboration is still higher for those who collaborate locally. Moreover, this learning-premium is increasing with time. Taken together, these two trends sketch a future in which collaborative learning in academia is dropping. As collaboration is a key platform over which knowledge flows (39, 40) and by which creativity is spurred (41), this could have severe consequences for the diffusion and production of novel knowledge in academia. Moreover, as existing knowledge is the key building block for generating novel knowledge and know-how, the overall rates of knowledge production might drop. Part of this phenomenon might be due to the internationalization of academia and the increasing pressure for scholars to specialize. In recent decades, academia has gone from a national enterprise geared towards national readerships to one with an international readership. Increased competition might favor the specialist over the generalist, encouraging platonic collaborations among specialists, each contributing their individual expertise. In such scenarios social learning is unlikely to happen or is reduced to a minimum.

The learning premium from local collaboration supports the arguments in the literature on localized learning, spatial clustering, and agglomeration in firms. In this literature, there are arguments for the various benefits for firms to cluster locally, ranging from the classical Marshallian externalities (e.g., reduced transportation costs, access to suppliers, etc.) to more intangible knowledge spillover. Geographical proximity is thought to facilitate repeated face-to-face interactions, fostering relationship building and allowing for rapid troubleshooting and serendipitous encounters, boosting opportunities for knowledge sharing and learning among employees of the firm. The evidence presented here suggests co-locating, albeit through the platform of formal collaboration, provides benefits at the micro-level for the individual scholar.

With subsequent significantly higher probabilities of acquiring know-how among scholars engaged in local collaborations, the stage is set for geographically localized *knowledge lock-in* with important implications for economic development. The localized nature of knowledge production is well described and characterized by spatially uneven distributions (13, 35, 42).

With the production of academic papers and patents predominantly occurring in relatively specialized (urban) locations, these locations would tend to benefit most from the economic spin-offs from the novel knowledge, such as new firm creation and certain first-mover advantages. The diffusion of new knowledge to other locations then becomes restricted, because knowledge flows are bounded geographically (43), although the precise mechanisms are still unclear (40). Thus, locations not producing the new knowledge would be unable to exploit it for job creation and other economic developments. Moreover, the locations producing new knowledge can combine it with other localized knowledge to produce more novel knowledge, leading to spatial inequalities of economic development. In this paper, we found that the diffusion of know-how through collaboration favors geographical proximity, and thus the flow of know-how across space is restricted. Indeed, with localized knowledge production, local diffusion and restricted flows across space, local accumulations of know-how and specialization occur.

Geographical localized knowledge lock-in has important implications for corporate innovation strategies and governmental economic development policies. The probability of success of corporate innovation strategies targeting fields of know-how that are outside the corporate know-how portfolio benefits from locating to where this know-how already exists. Access to know-how is one of the primary reasons why firms co-locate in high-rent locations such as the San Francisco Bay area. Our findings support these strategies, as the rates of acquiring know-how are higher when engaging in local collaboration. In addition, the evidence presented here suggest that governmental economic development policies should be hesitant to invest in non-local know-how, but rather should build upon local existing know-how. These ideas are well described in the literature on regional diversification (44), smart specialization (35), and product and technology space (21, 22, 45). Our research provides empirical evidence in support of these ideas at the micro-level.

The findings have two implications for the structuring of teams. If (one of) the objectives of team-work is to diffuse know-how across its members, limited team sizes and geographical proximity are beneficial. Individuals collaborating in smaller teams are more likely to acquire know-how than those collaborating in larger teams. This can be explained by the denser interactions allowed by smaller teams. More intense and trustful relationships can be cultivated in smaller teams to promote knowledge sharing, something that is less likely to happen in larger teams. Perhaps this is why smaller teams produce output that is more disruptive than the work of larger teams (46).

Materials and Methods

Data: The Microsoft Academic Graph (MAG) (23) includes over 60 million academic papers with over 200 million disambiguated (co-)authors as of January 1st 2018. Microsoft Research uses natural language processing algorithms to classify each paper into several 'Fields of Study' (FOS) based on their topics, fields, and methods. There are more than 200,000 different FOS across five hierarchies. Microsoft Research also ranks the affiliations of the authors of the papers based on prestige. The prestige of the affiliations is assigned to four evenly sized groups based on the quartile of the ranking distribution, with group 1 being the most prestigious.

Location of Authors: Authors were positioned in geographical space by geo-coding their first affiliation mentioned in a paper. We used the Google Maps API through the *ggmap* library (47) in Microsoft R Open 3.5.3 (48, 49). The name and full address of the affiliation was used to locate the affiliation and returns a set of coordinates. This allowed us to calculate the geographical distance between scholars. We assumed that an author's first listed affiliation corresponded to the author's primary place of residence at the time of publishing. Note that authors can be assigned to multiple locations throughout their career.

Sampling Strategy: We sampled 'focal scholars' with the following conditions: in a sequence of three consecutive papers authored by a scholar, the second in the sequence is a collaborative paper (number of co-authors > 1) and the third paper single-authored. The first paper in the sequence is used to observe the pre-collaboration knowledge portfolio of the scholar by recording all the FOS this scholar has published on. The third in the sequence needs to be a single-authored paper. These requirements allowed us to examine whether the focal scholar was able to produce a new FOS that she/he published on during the previous collaboration. This precludes any potential influence of the co-authors when examining if know-how is acquired.

Know-how Acquisition: By following a scholar's career over time, we can construct a cumulative knowledge portfolio that records all the FOS that the scholar is associated with at a certain period in time. Know-how acquisition occurs when the single-authored paper of the focal scholar (third paper in the sequence) contains at least one FOS that occurred in the collaborative paper (second paper in the sequence), but not in the pre-collaborative knowledge portfolio of the focal scholar (first paper in the sequence). This know-how is deemed to have been acquired during the collaboration process with the co-author(s) and/or directly from the co-author(s) (Figure 1).

Local Collaboration: Collaboration is considered local when the average geographical distance (as-the-crows-fly) between co-authors is equal or smaller than 700 m. This translates to a walking time of roughly 10 minutes. Increasing this boundary up to 1100 m did not quantitatively or qualitatively change the results of our analysis. Beyond a distance of 1100 m, the learning rates for local and non-local collaboration started to converge.

Learning Premium: The learning premium for a local collaboration is defined as the ratio of the learning rate of locally collaborating scholars divided by the learning rate of non-locally collaborating scholars minus one multiplied by a hundred. For example, a local learning rate of 10% and non-local rate of 5% gives ((10÷5) - 1) x 100 resulting in a learning premium of 100%.

Career Stage: A scholar is in their early career stage if their first and last publication were within 10 years. A scholar is in a mid-career stage if their first and last papers were published within 10 and 20 years. A scholar is considered to be in their late career stage if their first and last papers were published within more than 20 years.

Data Matching: We used a Coarsened Exact Matching (32–34) algorithm as defined in the *cem* library (50) in Microsoft R Open 3.5.3 (48, 49). We matched scholars who engaged in local collaborations with similar scholars in non-local collaborations by team size, affiliation ranking, discipline, year of collaboration, year of single-authored paper, career stage, and time difference between collaborative and single-authored papers. We dropped observations that were not matched. The resultant samples with locally and non-locally collaborating scholars looked similar in terms of co-variates. This allowed us to more precisely estimate the effect of distance (local or not-local) on acquiring know-how through collaboration, as the potential influence of confounding factors was reduced.

Statistical Model: A multivariate (linear) logistic regression model was constructed in Microsoft R Open 3.5.3 (48, 49) using the *att* function in the *cem* library (50). The model included team size and number of FOS from the collaborative paper, ranking of the affiliation (four groups), time difference (in months) between the collaborative and single-authored papers, career stage of the focal scholar, and fixed effects for the year of the collaborative and single-authored papers, and the discipline of the single-authored paper. The results of the full model are available in the SI.

Supervised Machine Learning Model: We trained and validated a supervised gradient boosting machine (GBM) model in Microsoft R Open 3.5.3 (48, 49) using the *h2o* library (see www.h2o.ai). We used this trained model to predict which observations, based

on the same set of predictors as in our main statistical model, have acquired know-how or not. The variable importance plot function *h2o.varimp_plot* was used to generate plots indicating how much each variable contributed to classifying an observation. The *DALEXtra* (51) and *iBreakdown* (52) were used to produce plots that showed the breakdown of how the trained GBM model makes a decision.

ACKNOWLEDGMENTS. We would like to thank our colleagues at the Kellogg School of Management, Northwestern Institute on Complex Systems and University of Hong Kong for their comments and feedback on earlier drafts of this paper. In particular, we would like to thank Brian Uzzi, Benjamin Jones, Dashun Wang, Olav Sorenson, David Rigby and Jason Owen-Smith for their feedback.

Bibliography

1. J Henrich, *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* (Princeton University Press), (2017).
2. R Boyd, PJ Richerson, J Henrich, The cultural niche: Why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci.* **108**, 10918 LP – 10925 (2011).
3. E von Hippel, Cooperation between Rivals: Informal Know-How Trading in *Indsutrial Dynamics*, ed. B Carlsson. (Springer Netherlands, Dordrecht), pp. 157–175 (1989).
4. DJ Teece, G Pisano, A Shuen, Management. *Strateg. Manag. J.* **18**, 509–533 (1997).
5. F Cairncross, *The death of distance: how the communications revolution will change our lives.* (Harvard Business School), (2001).
6. T Friedman, The world is flat. *New York: Farrar, Straus Giroux*, 488 (2005).
7. D Hummels, Transportation Costs and International Trade in the Second Era of Globalization. *J. Econ. Perspectives* **21**, 131–154 (2007).
8. GBTA, Business travel spending in the United States from 2011 to 2017 (in billion U.S. dollars), Technical report (2017).
9. R Hausmann, FMH Neffke, The workforce of pioneer plants: The role of worker mobility in the diffusion of industries. *Res. Policy* (2018).
10. M Storper, The Limits to Globalization : Technology Districts and International Trade. *Econ. Geogr.* **68**, 60–93 (1992).
11. EL Glaeser, HD Kallal, JA Scheinkman, A Shleifer, Growth in cities. *J. Polit. Econ.* **100**, 1126–1152 (1992).
12. R Florida, *Cities and the creative class.* (Routledge, London), (2005).
13. DB Audretsch, MP Feldman, R & D Spillovers and the Geography of Innovation and Production. *The Am. Econ. Rev.* **86**, 630–640 (1996).
14. K Morgan, The exaggerated death of geography: Learning, proximity and territorial innovation systems. *J. Econ. Geogr.* **4**, 3–21 (2004).
15. EE Leamer, M Storper, The economic geography of the internet age. *J. Int. Bus. Stud.* **32**, 641–665 (2001).
16. JW Sonn, M Storper, The increasing importance of geographical proximity in knowledge production: An analysis of US patent citations, 1975-1997. *Environ. Plan. A* **40**, 1020–1039 (2008).
17. M Polanyi, *Personal Knowledge: Towards A Post-Critical Philosophy.* (Routledge, London), (1958).
18. M Storper, AJ Venables, Buzz: Face-to-face contact and the urban economy. *J. Econ. Geogr.* **4**, 351–370 (2004).
19. DJ Teece, Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets. *California management review* **40**, 55–79 (1998).
20. P Maskell, A Malmberg, Localised learning and industrial competitiveness. , 167–185 (1999).
21. Ca Hidalgo, R Hausmann, The building blocks of economic complexity. *Proc. Natl. Acad. Sci. United States Am.* **106**, 10570–10575 (2009).
22. Ca Hidalgo, B Klinger, a L Barabási, R Hausmann, The product space conditions the development of nations. *Sci. (New York, N.Y.)* **317**, 482–7 (2007).
23. A Sinha, et al., An Overview of Microsoft Academic Service (MAS) and Applications in *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion.* (ACM Press, New York, New York, USA), pp. 243–246 (2015).
24. JG Foster, A Rzhetsky, JA Evans, Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
25. T Jia, D Wang, BK Szymanski, Quantifying patterns of research-interest evolution. *Nat. Hum. Behav.* **1**, 78 (2017).
26. TS Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Philosophy of science. (University of Chicago Press), (1977).
27. TS Kuhn, *The structure of scientific revolutions.* (University of Chicago press), (2012).
28. A Rzhetsky, JG Foster, IT Foster, JA Evans, Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci.* **112**, 14569–14574 (2015).
29. BF Jones, The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? *The Rev. Econ. Stud.* **76**, 283–317 (2009).
30. S Wuchty, BF Jones, B Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–9 (2007).
31. F van der Wouden, A history of collaboration in US invention: changing patterns of co-invention, complexity and geography. *Ind. Corp. Chang.* (2019).
32. DE Ho, K Imai, G King, EA Stuart, Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Polit. Analysis* **15**, 199–236 (2007).
33. SM Iacus, G King, G Porro, Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *J. Am. Stat. Assoc.* **106**, 345–361 (2011).
34. SM Iacus, G King, G Porro, Causal inference without balance checking: Coarsened exact matching. *Polit. analysis* **20**, 1–24 (2012).
35. PA Balland, D Rigby, The Geography of Complex Knowledge. *Econ. Geogr.* **93**, 1–23 (2017).
36. WM Cohen, DA Levinthal, Absorptive Capacity: A New Perspective on Learning and Innovation. *Adm. Sci. Q.* **35**, 128–152 (1990).
37. B Nooteboom, Learning by interaction: Absorptive capacity, cognitive distance and governance. *J. Manag. Gov.* **4**, 69–92 (2000).
38. R Boschma, Proximity and Innovation: A Critical Assessment. *Reg. Stud.* **39**, 61–74 (2005).
39. J Owen-Smith, WW Powell, Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organ. science* **15**, 5–21 (2004).
40. S Breschi, F Lissoni, Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *J. Econ. Geogr.* **9**, 439–468 (2009).
41. B Uzzi, J Spiro, Collaboration and Creativity: The Small World Problem. *Am. J. Sociol.* **111**, 447–504 (2005).
42. PA Balland, R Boschma, J Crespo, DL Rigby, Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Reg. Stud.* **53**, 1252–1268 (2019).
43. AB Jaffe, M Trajtenberg, R Henderson, Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Q. J. Econ.* **108**, 577–598 (1993).
44. F Neffke, M Henning, R Boschma, How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Econ. Geogr.* **87**, 237–265 (2011).
45. DL Rigby, WM Brown, Who benefits from agglomeration? *Reg. Stud.* **49**, 28–43 (2015).
46. L Wu, D Wang, JA Evans, Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).
47. D Kahle, H Wickham, ggmap: Spatial Visualization with ggplot2. *The R J.* **5**, 144–161 (2013).
48. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2019).
49. Microsoft, RC Team, *Microsoft R Open* (Microsoft, Redmond, Washington), (2019).
50. Iacus, et al., *cem: Coarsened Exact Matching*, (2018).
51. S Maksymiuk, P Biecek, *DALEXtra: Extension for 'DALEX' Package*, (2019).
52. A Gosiewska, P Biecek, iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models (2019).

Supplementary Information

There are differences in the likelihood of collaboration across academic disciplines. Figure 8 shows the ratio of single-authored to co-authored papers across all papers by disciplines in the Microsoft Academic Graph (MAG) snapshot of as of January 1st 2018. Disciplines below the red dotted horizontal line are characterized by teamwork rather that single authorship. These disciplines tend to be associated with STEM fields. The disciplines above the red dotted line are characterized by single-authored publications and tend to be associated with the social sciences. The differences in the tendency to collaborate across disciplines might confound the relationship between geographical distance and acquiring know-how during collaboration. For example, in fields with a smaller number of possible collaborators, there might be more competition and stronger selection effects, resulting in the selection of only the very best scholars for collaboration. This could boost overall learning rates, regardless of geographical distance. Another example is the difference in the spatial distributions of a discipline. Disciplines that cluster in space are more likely to have greater shares of local collaboration, with everything being equal. The effect of distance on learning might be different among scholars in these disciplines compared to those working in other disciplines. For these reasons, we included discipline-fixed effects in our statistical models.

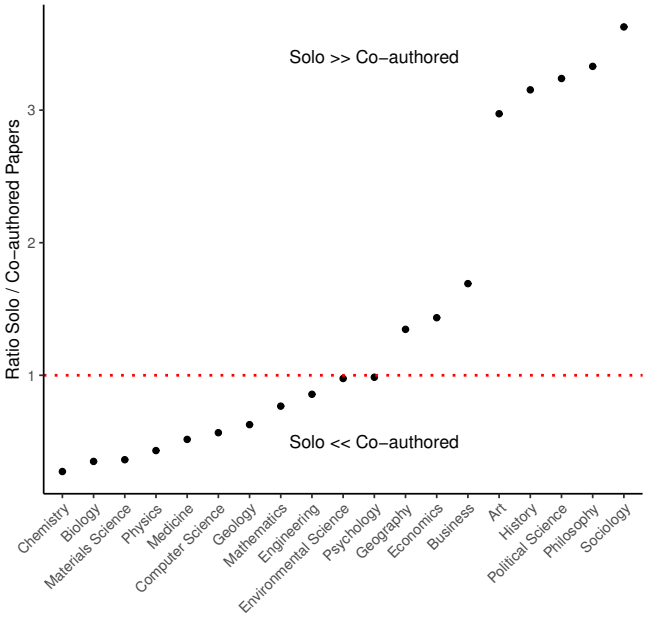


Fig. 8. Ratio of solo over co-authored papers by discipline.

The negative impact of geographical distance on acquiring know-how was observed across all academic disciplines, but differed in strength by discipline. Figure 9 shows the impact of increasing geographical distance on acquiring know-how through collaboration associated with academic disciplines, with STEM-type fields being hit the hardest. These results are obtained using a simple logistic regression model, in which acquiring know-how is regressed onto geographical distance in meters for each academic discipline, with the intercept set to zero to compare the slopes. Note that we do not control for any co-variates in this model, as we did in Table 2d.

In Figure 10, the learning-premium from local collaboration along the 15 years of data is plotted for the 1980, 1990, and 2000 cohorts. This figure differs from that in Figure 2C, in that the slopes of each cohort were estimated from the first 15 years of data for all cohorts. This allows us to compare the slope of the different cohorts and examine whether there are meaningful differences between the cohorts. The slopes of these cohorts look qualitatively similar, suggesting the learning premium from local collaboration does not change over time. This supports our main findings indicating that

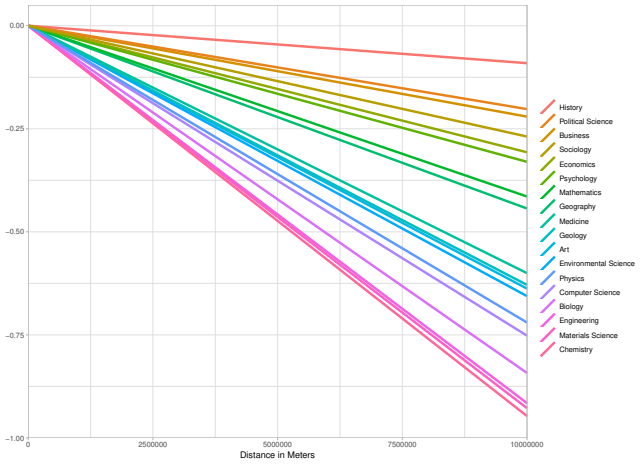


Fig. 9. Probability of learning by local and non-local collaboration.

geographical proximity matters for acquiring know-how, which was not affected by technological advances introduced between 1980 and 2000.

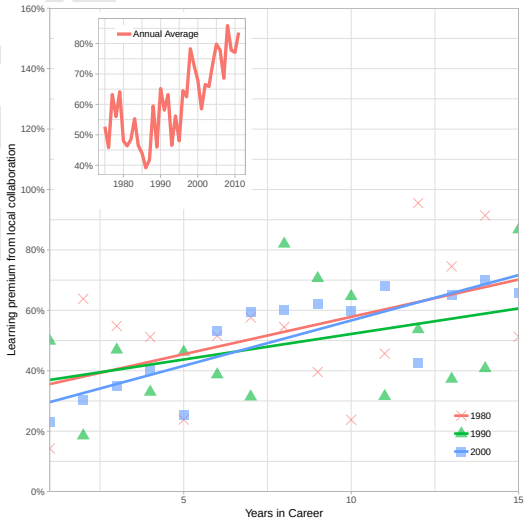


Fig. 10. Learning premium from local collaboration for the first 15 years of data by the 1980, 1990 and 2000 cohort.

The statistical model of the matched data estimated the linear effects of predictors on acquisition of know-how, with possible complex interactions or non-linear relationships not accounted for. To examine whether our results hold under such conditions, we constructed a supervised machine learning classification model. Observations from our sample were randomly assigned to a training set (80%), validation set (10%) or test set (10%). We trained and validated a supervised gradient boosting machine (GBM) model. This trained model was used to predict which observations, based on the same set of predictors (plus affiliation of single-authored paper) as in the statistical model, have acquired know-how or not. Our trained model had an accuracy of 76% on the out-of-sample test set. Figure 11A shows the relative importance of the variables in the model used to make the predictions. This figure indicates that the distinction between local and non-local collaboration could be used to predict acquiring know-how. This was found to have more influence on the classification than discipline, rank, year of single-authored paper, and affiliation. Figure 11B and 11C provide the breakdown of how the trained model makes decisions for local and non-local collaborations, respectively. Local collaboration contributed to the prediction of acquiring know-how,

whereas non-local collaboration had a negative effect on acquiring know-how. Note that the cut-off point in our GBM model was an optimized F1 score of 0.11, not the usual 0.5 or 50%. Evidence from this model suggests that local collaboration promotes know-how acquisition among scholars.

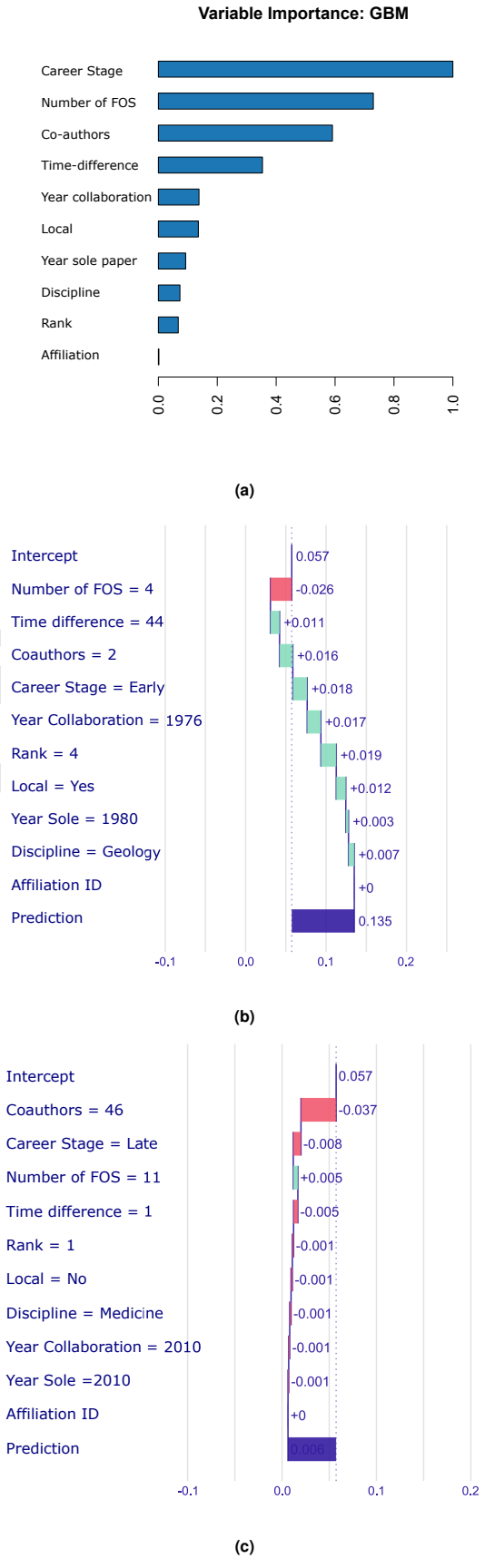


Fig. 11. Supervised machine learning model: (a) Variable importance of a trained model for predicting know-how acquisition through collaborating and examples of the breakdown of the classification for a case of (b) local and (c) non-local collaboration.