LSE International Inequalities Institute

WORKING PAPER 103

The transformative effects of tacit technological knowledge

Sergio Petralia
Department of Economic Geography, Utrecht University

Tom Kemeny
Munk School of Global Affairs & Public Policy, University of Toronto; International Inequalities Institute, LSE

Michael Storper
Department of Geography and Environment, LSE; University of California, Los Angeles

SEPTEMBER 2023

Sergio Petralia
Department of Economic Geography, Utrecht University

Tom Kemeny
Munk School of Global Affairs & Public Policy, University of Toronto; International Inequalities Institute, LSE

Michael Storper
Department of Geography and Environment, LSE; University of California, Los Angeles

In addition to our working papers series all these publications are available to download free from our website: www.lse.ac.uk/III

For further information on the work of the Institute, please contact the Institute Manager, Liza Ryan at e.ryan@lse.ac.uk

International Inequalities Institute
The London School of Economics
and Political Science, Houghton Street,
London WC2A 2AE

E Inequalities.institute@lse.ac.uk
W www.lse.ac.uk/III
@LSEInequalities

LSE International Inequalities Institute

The transformative effects of tacit technological knowledge

Sergio Petralia*†, Tom Kemeny††, and Michael Storper†††

†Utrecht University, Department of Economic Geography
††University of Toronto, Munk School of Global Affairs & Public Policy
†††University of California, Los Angeles & London School of Economics

June 22, 2023

Abstract

Tacit knowledge – ideas that cannot readily be meaningfully and completely communicated – has long been considered a precursor to scientific and technological advances. Using words and phrases found in the universe of USPTO patents 1940-2020, we propose a new method of measuring tacit knowledge and its progressive codification. We uncover a discontinuity in the production of highly tacit technologies. Before 1980, highly- and less-tacit inventions are evenly distributed among inventors, organizations, scientific domains and subnational regions. After 1980, inventors of highly tacit patents become relatively rare, and increasingly concentrated in domains and locations. The economic payoffs to tacit knowledge also change, as it starts unequally rewarding high-income workers. This suggests a role for tacit knowledge in contributing to the rise in income inequality since 1980.

*Corresponding author. Email: s.g.petralia@uu.nl.

1 Introduction

Over the last two centuries, humanity has experienced dramatic improvements in living standards and economic growth, fueled in large part by the widespread application of new scientific and technological knowledge to human activity (Mokyr, 1992; Landes, 2003; Maddison, 2007). A key element in this process is how sparks of human ingenuity, initially accessible only to a select few 'in the know,' become available to be used by a wider community. This transition – from emergent, unstable ideas to widely-understood principles – can be characterized as a shift between tacit and increasingly codified knowledge.

Tacit knowledge refers to ideas that cannot readily be communicated in a manner that is "meaningful and complete" (Teece, 1998). Not all tacit knowledge is new, but new technologies tend to be rich in tacit ideas, because their meaning has not yet been widely agreed upon, and they lack standardized codes that facilitate communication. In the words of the 20th century chemist and philosopher Michael Polanyi, individuals with new tacit ideas "know more than they can tell" (Polanyi, 1966). Note that knowledge in science and technology is tacit not when a particular individual cannot transmit it, but rather when communities have not yet established widely accepted codes that describe it. Though such tacit knowledge can be shared, dissemination proceeds only through intensive interpersonal interaction – a requirement that strongly limits its diffusion. As a result scientific and technological fields are marked by a strong push to codify tacit knowledge.[1] Progressive codification allows a wider community of users to apply, extend and recombine new ideas, which in turn drives scientific, technological and economic advances (Romer, 1990; Solow, 1956; Mokyr, 2009).

Despite longstanding scientific interest and societal importance, our understanding of tacit knowledge and its codification remains mostly conceptual and anecdotal (Polanyi, 1966; Teece, 1998; Saviotti, 1998; Lam, 2000). In this article we propose a new measure of tacit knowledge, focusing on ideas linked to scientific and technological progress. Taking a 'Science of Science' approach (Fortunato et al., 2018), we use the texts of patent documents to identify technical words and concepts that exhibit rapid growth in appearances across many patents. We call such instances of rapid proliferation, 'rushes of codification.' These rushes are palimpsests of ideas that were highly tacit only moments before they are described in a patent application. Their proliferation across the patent system signals growing centrality in domains of scientific knowledge.

There are many reasons to be interested in the transition from tacit to increasingly codified knowledge. Rapidly codifying knowledge is thought to play a vital role in shaping the wider process of technological change (Nelson and Winter, 2002; Saxenian, 1996). Inventions rooted in rapidly codifying knowledge indicate the rate and direction of productivity changes , as well as

[1] This is to be contrasted against some other domains of human activity, such as art or music, where the push to codify tacit knowledge is less strong and less central.

how organizations and workers reap pecuniary rewards from innovation (Autor, 2014). Tracking the emergence of tacit knowledge and its codification also sheds light on current debates about science and technology, including whether and why new, disruptive ideas are getting harder to find (Park et al., 2023; Bloom et al., 2020); if innovation is increasingly less subject to competition (Aghion et al., 2005); and the extent to which recent waves of technological change are exacerbating inequality between people and places (Frank et al., 2019; Balland et al., 2020; Kemeny et al., 2022).

In order to build our measure of tacit knowledge, we analyze detailed textual descriptions of every patent document granted by the U.S. Patent and Trademark Office (USPTO) for the century spanning 1920 to 2020. We track the emergence of technical words and phrases as they enter the lexicon, and isolate the subset exhibiting rushes of codification. We then identify individual patents in which these concepts are concentrated. We link patent-level information to a new dataset that, for nearly all patents over our study period, documents the names of inventors and their organizations (See SOM, Section A for details). This enables us to explore who produces highly tacit knowledge, the organizations in which they belong, the cities in which they reside, and domains of knowledge to which they contribute. We also explore the connections between tacit patenting and local income distributions.

2 Using patent documents to identify tacit knowledge

To illustrate the transition from tacit to increasingly codified knowledge, consider Alexander Graham Bell's invention of the telephone. In the 1870s, Bell aimed to create a viable multiple telegraph, an electrical device in which the same wire is used to transmit multiple text messages between two points. Bell's novel solution was called the induction telephone. This device rested on ideas that arose out of his speech-therapy work with children, but also incorporated concepts from a system of tuning forks developed by the German scientist Hermann von Helmholtz, as well as discoveries Bell made around the conversion of sound waves into fluctuating current (Gorman and Carlson, 1990; Bell, 1876). While his invention leveraged preexisting and new ideas, his selection of which pieces of knowledge to use and how to recombine them were unique and embodied – a function of his particular mind and lived experiences. Prior to patenting, these notions existed chiefly in Bell's mind – they were tacitly held. The few people who knew that this kind of assemblage might be possible, like his assistant Thomas Watson, had to learn about it through direct interactions with Bell.

Through the text and diagrams included in his initial patents (Bell, 1877, 1878), new concepts were traced with sufficient detail and clarity as to allow others to reproduce Bell's device. Patents thus placed these concepts in the public domain, enabling a flurry of inventive activity building on Bell's idea – better telephones, and other complementary innovations in the emergent areas needed to build a viable telephone industry. We can think of this flurry as formerly tacit knowledge being rapidly codified. By contrast, a century and a half later, we should expect there to be only a relatively modest amount of translation from tacit to codified knowledge around Bell's ideas, as the work of codification in that area was largely achieved long ago.

Generalizing from this example, the granting of a patent not only marks the provision of a new intellectual property right, it also signals a moment at which highly tacit technological concepts begin a journey of codification. This journey may never fully end, and over time may proceed at varying speeds. We measure aspects of this transition by exploiting detailed textual descriptions of around 9.2 million utility patent documents – each such patent granted by the U.S. Patent and Trademark Office (USPTO) between 1920 and 2020. We track new technical words and phrases as they enter the lexicon, taking their initial appearance and subsequent mentions in patent documents to cast light on the movement from tacit to codified knowledge.

2.1 Operationalizing tacitness

The patent system is a rich repository of codifying technological knowledge. In order to be granted a patent, inventors must transparently demonstrate the novelty and non-obviousness of the invention. Patents are costly to obtain, and hence should be limited to technological ideas with potential economic effects. Patent documents also contain a range of information about inventors and assignees[2], including their geographical location.[3] Using addresses and organizational information, we can consider the evolution of tacit knowledge not just in time but also in space, across organizations, and between technological domains.

We start by organizing knowledge contained in the patent system into a formal hierarchical structure. In this structure, a *concept* represents the smallest unit of technological knowledge, indicating a single idea that can be undergoing codification to a greater or lesser degree. At its most micro level, a single word can represent a concept. An *invention* is then a collection of concepts, linked together in such a way that creates a unique or novel device, method, composition or process. An invention can then be categorized through shared common principles or features into one or more *technological domains*.

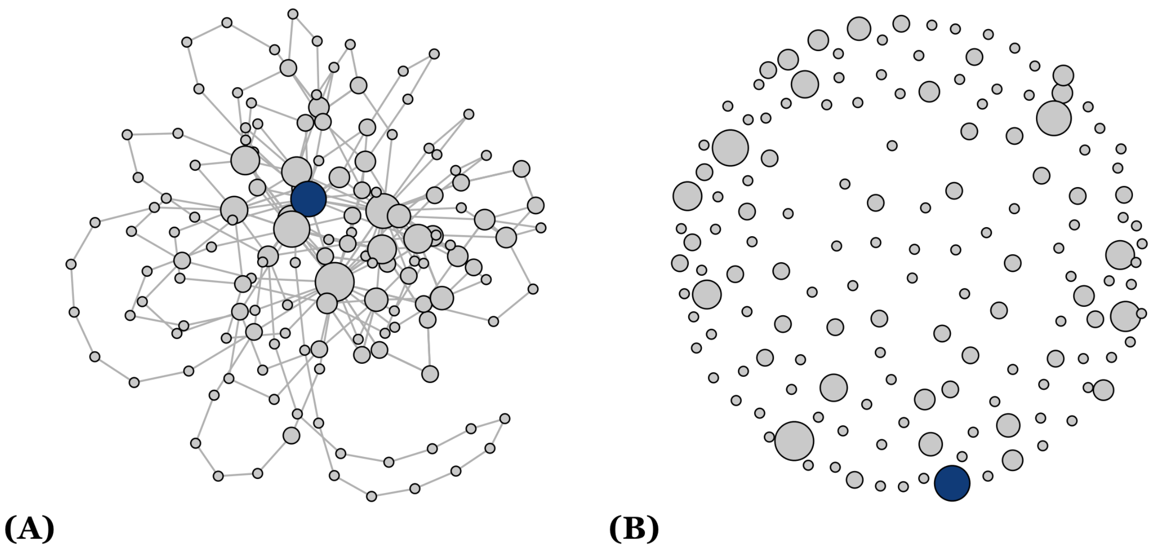
We represent the human knowledge found in patent documents as a network. Each node of the network represents a concept with at least some degree of codification, and each edge is a relationship between concepts. At a given moment in time, the sum of these components represents

[2] Information about inventors' and assignees' names after 1975 can be obtained already in a structured form at https://patentsview.org/download/data-download-tables. Prior to 1975 we extracted inventors' and assignees' names from digital documents by the 'Annual report of the Commissioner of Patents' and the 'Index of patents issued from the United States Patent Office' both available at the Smithsonian Libraries (https://library.si.edu) and from digital versions of the original patents provided by the USPTO at https://bulkdata.uspto.gov. See SOM Section A for more details about the data extraction process from these digital documents.

[3] Information about inventors' and assignees' geographical locations after 1975 can be obtained already in a structured form at https://patentsview.org/download/data-download-tables. For the Period prior 1975 we rely on (Petralia et al., 2016b,a).

the state of codified collective technological knowledge contained in patents. A new invention can be defined as a series of edges between nodes, some of them new, which integrates the invention with the existing state of codified technological knowledge. This integration or addition redefines the existing state of technological knowledge by adding new meanings or relationships (strengthening ties or edges) between concepts (nodes).

Figure 1: Semantic network of USPTO Patent #201,488: A.G. Bell's Speaking Telephone



Note: This graph provides two different network interpretations of Alexander Graham Bell's seminal telephone patent in 1878. Panel (A) displays a network representation in which words in the patent are connected to each other by the order in which they were written in the process of codifying the idea. Panel (B) considers that all words in the patent document are related to each other to some extent (we do not draw lines here, otherwise they would overtake the graph). The blue dot represents the word 'telephone'.

A primary challenge we face is to determine how words are related in a document to produce meaning, context, and ultimately, codification. To illustrate such a structure on a micro-scale, consider Figure 1, which offers a network representation of Alexander Graham Bell's initial telephone patent, with a node for each technical word.[4] The figure displays two of many potentially valid modes of semantic meaning-making around the central idea of the 'telephone' (in each panel, this word appears in blue). In Panel (A), words are connected to each other by the order in which they were written. In this view, terms that come later in a sentence are used to explain and provide context to those found earlier. Panel (B) starts from a different premise: we make no prior semantic assumptions, and thus consider all words in the document to be equally interrelated. Hence,

[4] https://patentimages.storage.googleapis.com/d1/80/dd/d765ce0184fcf3/US201488.pdf

each node is connected to each of the others (though we do not draw these lines, as they would overwhelm the figure). One could also motivate Panel (B) by assuming that we cannot identify which specific words are the main recipients of the context being provided in a given document's codification process, so we may as well consider all possibilities.

How can we determine which kind of semantic structure is at work in Bell's patent? More generally, how can we determine how meaning is generated not just in a single patent document, but for millions of them? We propose a heuristic solution, based on the frequency with which individual technical words and word pairs (bigrams) appear across patent documents. A word that does not appear frequently in patent texts ought not to be well-integrated into, or relevant for, the larger body of collective knowledge. By contrast, if a word experiences rapid growth in the number of times it appears across patent texts, this frequent re-use suggests that it is becoming a convenient shorthand, calling up agreed-upon meanings in the technological field at hand. A rapid burst of mentions is thus a trace element of a process where such meanings are being established, confirmed, and diffused. At more mature phases of this process, users share an efficient way of referring to meanings, facilitating more precise and wider communication. This in turn enables closer integration into collective scientific knowledge, without the need for intensive interpersonal contact among a small group of initiates.

The heuristic we propose is consistent with different theories of semantic meaning-making. For instance, in Panel (A) of Figure 1 network centrality could effectively identify concepts that are receiving meaning (Vega-Oliveros et al., 2019); thus, the word 'telephone' is central in this network. However, note that a measure of network centrality of words such as the degree centrality will have a one-to-one correspondence with the frequency with which words appear. This is because there is only one directional link for every concept in the document. The same argument can be made regarding the relationship between the frequency with which concepts appear and the degree distribution of the network described in panel (B), provided that links are weighted accordingly. Therefore our heuristic fits well across many possible meaning-making relationships between concepts, making frequency counts a reasonable starting point.

To operationalize these ideas, we generate a dictionary of technical concepts that describes the contents of all patents granted between 1920 and 2020. We start by identifying each unique word and bigram that appears across all patents. After a series of text cleaning routines, detailed in SOM Section B, the resulting dictionary contains 210,491 distinct concepts.

To measure the timing and extent with which a concept has experienced a rush of codification we consider two observable features of the distribution of words and bigrams. First, assuming that concepts exhibiting a rapid increase in mentions across the universe of patent texts are likely to be undergoing a process of codification, we define K as the knowledge codified around each concept w up to year t as:

$$K_{w,t} = \sum_{i=1920}^{t} \sum_{d} O_{w,d,i}$$

where $O_{w,t}$ represents the number of occurrences of concept w across patent documents d, from 1920 up to year t. We then identify concepts that experience rapid growth in K based on the relative speed at which they accumulate knowledge around them, measured as $\Delta Q(K_{w,t})$, where $Q(.)$ represents the quantile each concept occupies in the distribution of mentions until year t and Δ represents its growth. Though all concepts are continuously transformed through the addition of new and reshaped relationships, some experience this transformation more rapidly.

For intuition, consider Figure 2, which, across the corpus of patent texts, tracks how selected concepts move through the quantile distribution of mentions $Q(K_{w,t})$. The concept 'telegraphy' ranks high in the distribution, signalling its overall importance; however, over the study period, it is highly stable or declining in importance. We interpret this to mean that, though many new inventions are appearing that are reliant on this concept, telegraphy is not experiencing a rush of codification. Instead, by the time the study period begins, this concept is already highly codified. Meanwhile, Figure 2 reveals rushes of codification for 'transistor,' 'capacitor,' and 'internet', each at different periods of time. Each concept rises rapidly through the distribution of mentions in patent texts in a particular period, and then subsequently stabilizes. More recently, 'smart contract', a term specific to blockchain technologies, grows rapidly in the distribution of mentions. These patterns confirm the usefulnees of this approach in detecting the progressive codification of key terms in periods of intensive technological development.

Secondly, we define the breadth of mentions, B, as:

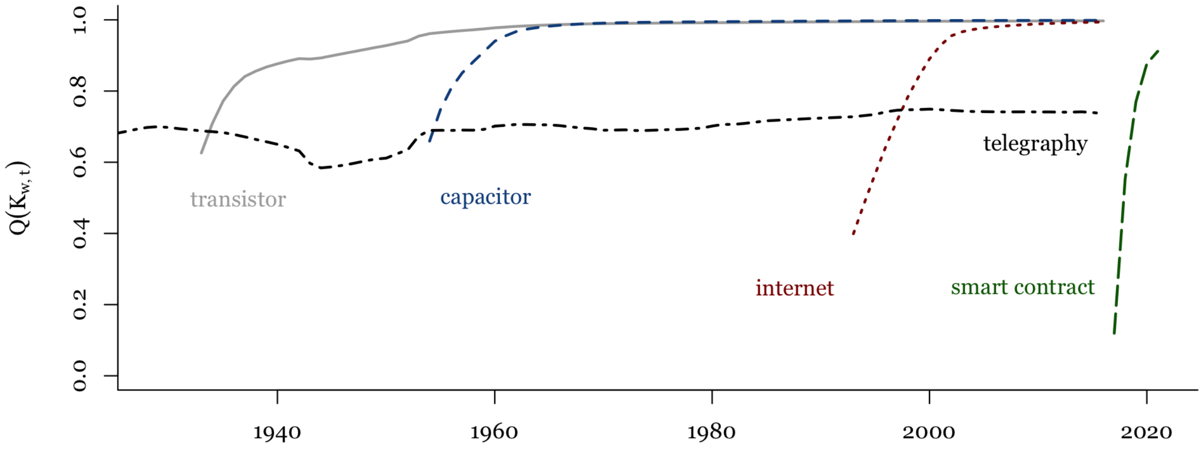
$$B_{w,t} = \sum_{d} max(I[O_{w,d,t} > 0])$$

where the indicator function $I[.]$ takes a value equal to one whenever a certain concept was used in a document. Therefore, $B_{w,t}$ captures the quantity of documents in which concept w appears in a given year (t). As before, we measure relative document occurrences using the quantile each concept occupies in a given year. In practical terms, this measure helps us distinguish redundant from non-redundant codification, since certain patents repeatedly mention particular concepts.

Using these two measures, we can identify concepts undergoing rushes of codification as those that are at or above certain percentile threshold in the distribution of $Q(\Delta K_{w,t})$ and $Q(B_{w,t})$ across all concepts in a given year. As an example, in 2020, the term 'smart contract' lies at the 99th percentile in terms of of $Q(\Delta K_{w,t})$, and at the 87th percentile in terms of $Q(B_{w,t})$. In this article we fix this threshold to the top 25%, although in the SOM Section D.3 we show that our results are not materially sensitive to changes to it.

The next step in the analysis is to aggregate from the level of an individual concept to that

Figure 2: The codification of concepts over time



Note: This graph shows the evolution over time of different words in the entire distribution of technological concepts. We show the percentile occupied by each word in each year, based on the value of $Q(K_{w,t})$.

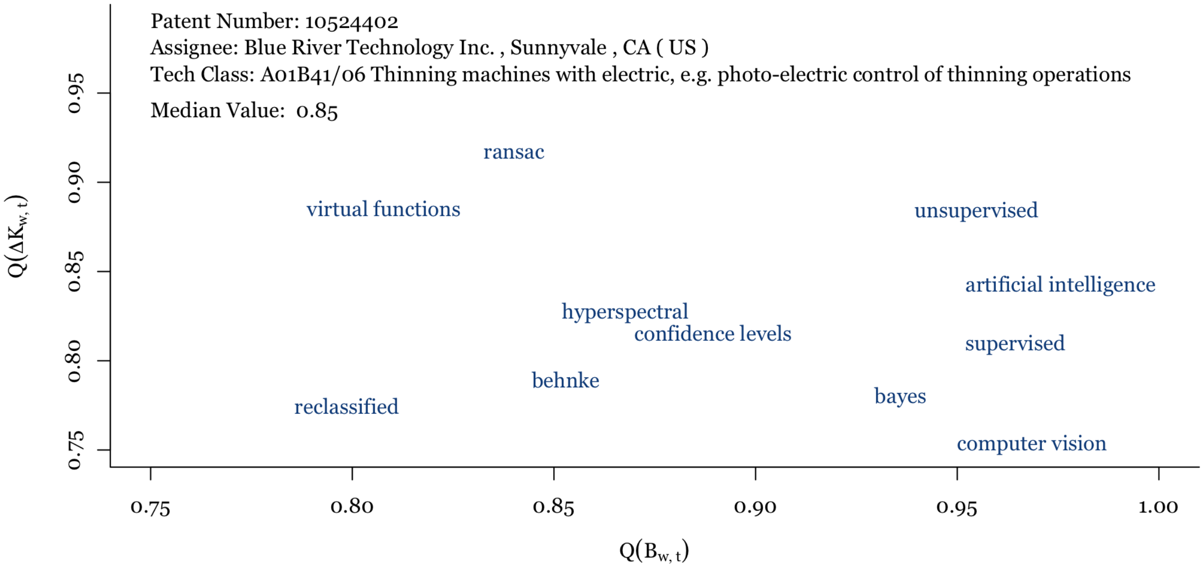
of an invention (patent) that can be identified as more or less tacit. To achieve this, we identify each concept in each patent that is undergoing a rush of codification, and take the median value of $Q(\Delta K_{w,t})$ and $B_{w,t}$ across all of these concepts.[5] A patent with concepts undergoing rapid rushes of codification will be ranked higher than another whose 'rushes' are more anemic. Meanwhile, patents containing no concepts undergoing such rushes will be considered the least tacit.

Figure 3 shows an example of how we calculate the relative ranking of an individual patent. It describes the contents of patent #10,524,402, the first to be granted in 2020. The patent, which incorporates machine learning techniques into agricultural machinery, contains several concepts that are highly tacit in that year, like 'unsupervised' or 'bayes'. Taking the median value of $Q(K_{w,t})$ and $B_{w,t}$ across all of these concepts yields a score for patent #10,524,402 of 0.85.

To simplify the analysis to follow, we dichotomize the distribution of patents in each year, assigning those with rankings above the median to the class of 'highly tacit' patents, with those below deemed 'less tacit.' Splitting the distribution in half in this way is a deliberately conservative gesture – it is of course impossible to pinpoint with certainty a threshold beyond which patents

[5] We limit our attention only to concepts undergoing rushes of codification based on the idea that a patent containing highly tacit knowledge can have only a few tacit words surrounded by a larger number of words that are already relatively codified. Practically, in Bell's initial patent, the tacit concept of the 'telephone' must be defined in terms of ideas that are already relatively well-understood, or codified. Taking the median value across *all* words in a patent could produce a mis-classification error, whereby patents like Bell's would in its day be incorrectly identified as being relatively well-codified.

Figure 3: Evaluating the tacitness of USPTO Patent #10,524,402



Note: This figure displays all tacit words found in patent #10,524,402. The horizontal axis shows the relative position of the words in terms of the amount of different contexts (documents) they were mentioned. The vertical axis shows the relative speed at which concepts were codified in 2020.

are or are not highly tacit, but with some confidence we can distinguish more- from less-tacit. In the SOM Section D.4 we show that results do not change significantly when a different threshold is considered.

For descriptive purposes, consider Table **??**. For the year 2020, it lists the ten most and least highly tacit technological domains, ranked according to the proportion of their patents which we describe as 'highly tacit.' To group patents into domains, we use class definitions from the Cooperative Patent Classification (CPC) system. In Table **??**, the most highly tacit patent classes lie mostly within the Computer & Communications category, in line with recent contributions around 'disruptive' innovation (Bloom et al., 2021), which identify the recent emergence of data science, artificial intelligence, and blockchain technologies. At the bottom of this ranking we see technologies that can be considered at a mature stage of development, including 'Rolling of Metal' and railway-related technologies. This descriptive evidence suggests a broad consistency between intuition and the results produced using our measure.[6] Further, class size appears to have little

[6] The appearance of technologies like 'HARVESTING; MOWING' in the top of 2020 is not a mistake, actually we are identifying concepts undergoing rushes of codification that are nonetheless found in longstanding domains. The example shown in Figure 3 falls into this category.

Table 1: Most and least highly tacit technologies, 2020

CPC Code	Share Highly Tacit	Total Patents	Class Name
G06N	0.424	1,878	COMPUTING ARRANGEMENTS BASED ON SPECIFIC COMPUTATIONAL MODELS
G01W	0.343	67	METEOROLOGY
G05D	0.322	1,527	SYSTEMS FOR CONTROLLING OR REGULATING NON-ELECTRIC VARIABLES
H04L	0.307	21,307	TRANSMISSION OF DIGITAL INFORMATION, E.G. TELEGRAPHIC COMMUNICATION
A01D	0.304	418	HARVESTING; MOWING
G06Q	0.301	7,993	DATA PROCESSING SYSTEMS OR METHODS...
G10L	0.300	1,905	SPEECH ANALYSIS OR SYNTHESIS; SPEECH RECOGNITION; SPEECH OR VOICE PROCESSING..
G06F	0.291	36,067	ELECTRIC DIGITAL DATA PROCESSING
B60D	0.290	107	VEHICLE CONNECTIONS
G06T	0.286	6,090	IMAGE DATA PROCESSING OR GENERATION, IN GENERAL
...	...	...	...
A61P	0.010	193	SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS
G03G	0.012	1,904	ELECTROGRAPHY; ELECTROPHOTOGRAPHY; MAGNETOGRAPHY
A23C	0.012	81	DAIRY PRODUCTS, E.G. MILK, BUTTER OR CHEESE; MILK OR CHEESE SUBSTITUTES; MAKING THEREOF
D21F	0.019	52	PAPER-MAKING MACHINES; METHODS OF PRODUCING PAPER THEREON
A01H	0.024	1,047	NEW PLANTS OR ; NON-TRANSGENIC; PROCESSES FOR OBTAINING THEM...
C07J	0.027	74	STEROIDS
B21B	0.027	73	ROLLING OF METAL
D01F	0.032	94	CHEMICAL FEATURES IN THE MANUFACTURE OF ARTIFICIAL FILAMENTS, THREADS, FIBRES...
C07K	0.032	2,350	PEPTIDES
B61F	0.034	59	RAIL VEHICLE SUSPENSIONS
C07D	0.035	2,878	HETEROCYCLIC COMPOUNDS

Note: This table lists Cooperative Patent Classification (CPC) classes, divided into the ten ranked as the most highly tacit (top half of the table) and least highly tacit (bottom half of the table. This ranking is achieved on the basis of the share of USPTO patents in each class in 2020 that is deemed to be highly tacit – above the median level of tacitness in a given year. See main text for definitions of tacitness at the level of individual concepts, as well as patents. Class names are abbreviated for fit.

relationship to tacitness, with larger and smaller categories featuring equally across the most- and least-tacit. An analogous table for 1930 in SOM Section B.5 reveals patterns for that period that concord with a vast historical literature documenting this period (David, 1990; Lipsey et al., 2005; Bresnahan and Trajtenberg, 1995; Field, 2011; Moser and Nicholas, 2004), with highly tacit technologies emerging from electrical & electronic, chemical, and combustion engine technologies, and mature mechanical and textile technologies appearing as least tacit.

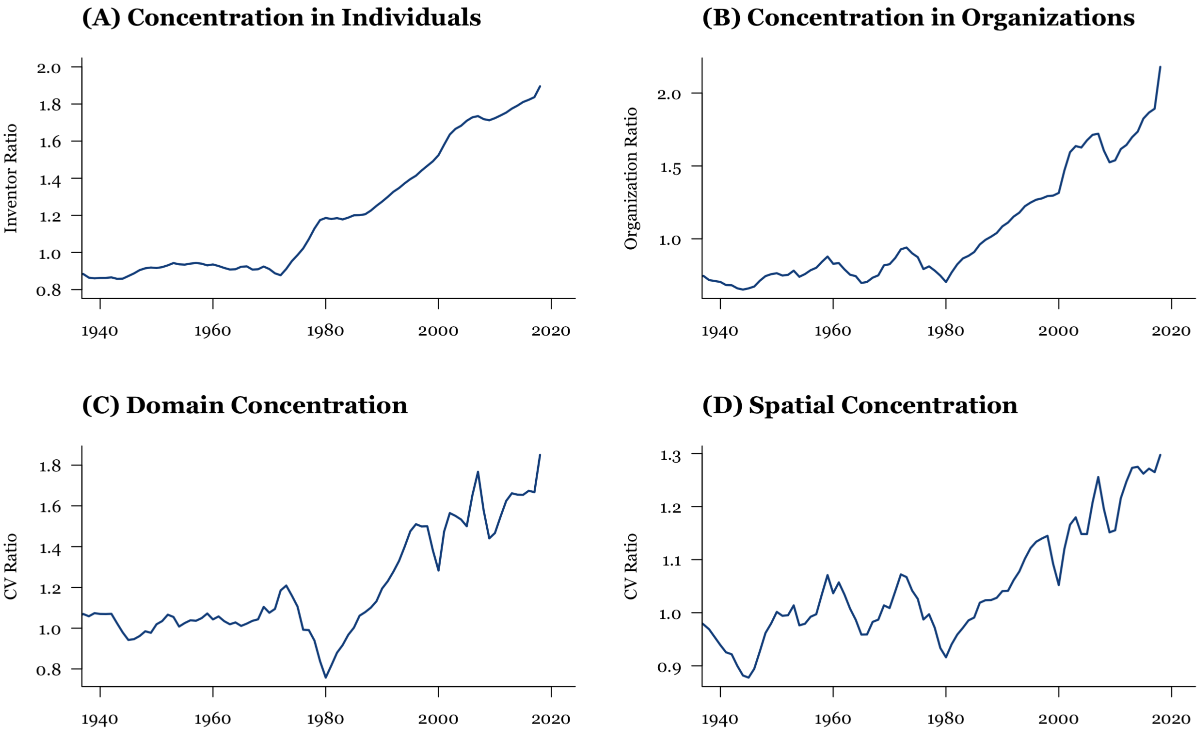
3 The Changing Face of Tacit Knowledge Production

We now explore whether and how the distribution of tacit knowledge in the United States has changed over the 80-year study period. We consider its distribution across four dimensions: individual inventors; inventing organizations; technological domains; and subnational locations. Building on a long tradition of studies examining the links between technology and economic inequality, (Katz and Murphy, 1992; Autor et al., 2008a; Goldin and Katz, 2009) we then exploring the association between highly tacit patents and the distribution of income.

Figure 4 reveals changes in how patents containing highly tacit technological knowledge – patents with tacitness values that place them in the top half of all patents granted in a given year – are distributed relative to less highly tacit patents – those in the bottom half of the distribution.

Panel (A) explores this distribution among inventors. For a given year, the y-axis of this chart captures the ratio of the number of individuals producing relatively non-tacit inventions to the number producing highly tacit inventions. We consider an inventor to be a producer of highly tacit

Figure 4: Tacitness in people, organizations, technological domains, and American cities, 1940–2020



Note: Panels (A) and (B) track the annual ratio of the number of inventors producing less- to more-highly tacit USPTO patents, with the median tacitness rank dividing patents in a given year into two equal-sized groups. Panels (C) and (D) capture changes in the annual ratio between the coefficients of variation for more- and less-highly tacit patents. Higher values of this ratio indicate greater concentration in either Cooperative Patent Classification (CPC) classes (domains of technological knowledge) or 1990-vintage Commuting Zones (regions) for more highly tacit patents as compared with less tacit patents.

patents in year t if they are listed on at least one patent with a tacitness score above the median; in each year, categories of inventors are thereby mutually exclusive. Given that, by construction, half the patents in each year are placed into either more- or less-tacit categories, the naive expectation is that approximately half the inventors will fall into either group. This appears to be approximately true over the first forty years of the study period: the ratio of more- to less-tacit inventors hovers just below one (around 53% of inventors produce at least one highly tacit patent over the 1940 to 1980 period). Then, starting in the second half of the 1970s, the ratio suddenly begins to rise, signalling a growing relative scarcity of inventors in the patent system capable of producing highly tacit technological knowledge. Hence, despite having tacit patents in each year pegged at half of the total new patents granted, we observe a progressive reorientation of the inventive workforce in which tacit patents are the output of a narrowing share of inventors. From the late 1970s onward, as a subset of all inventors, those capable of producing tacit knowledge are becoming increasingly scarce. Note that this does not mean that absolutely fewer inventors are producing tacit knowledge. Indeed, the study period tracks a major expansion of patenting, driven by a growing global community of inventors. But it points to increasing concentration in which individuals produce tacit technological knowledge, such that, by 2020, individuals are almost twice as likely to produce a less-tacit than more-tacit invention.

Panel (B) of Figure 4 considers a similar ratio, this time tracking the distribution of tacit patents across organizations. Patterns in this figure closely mirror those shown in Panel (A). We observe a pattern of stability and approximate equality over the initial 40 years of the study period; in each year between 1940 and 1980, organizations producing highly tacit patents made up between 55 and 60 percent of all organizations that were granted patents in a given year. Then, around 1980, the share of organizations producing tacit patents starts to fall, and continues to fall thereafter. Thus, in the first half of the study period, a randomly selected patenting organization was slightly more likely to have produced a highly tacit patent; by 2020 they are more than twice as likely to be producing less-tacit patents.

Figure 4 Panel (C) considers the concentration of tacit technologies across domains of scientific and technological knowledge, as captured in the Cooperative Patent Classification (CPC) system. The y-axis of the Panel (C) captures the ratio of the coefficient of variation for highly tacit over less tacit patents across patent classes. Higher values of each coefficient of variation point to greater domain concentration. A ratio between them equal to one would indicate that more and less tacit patents are equally distributed across the different domains of scientific and technological knowledge, whereas a value above one means that highly tacit patents are more concentrated in a narrower range of classes. Interpreting Figure 4 Panel (C), between 1940 and 1980, more and less tacit patents were both relatively widely distributed across technological domains, with tacit patents only slightly more concentrated. From 1980, this pattern changes dramatically. Highly tacit patents began concentrating in a limited subset of classes, indicating a greater narrowing in the kinds of technologies from which tacit inventions are emerging.[7] Though there is some year-on-year noise, this narrowing trend continues throughout the post-1980 period, such that by 2020, highly tacit inventions are almost twice as concentrated as less tacit patents.

Panel (D) of Figure 4 tracks the subnational geographical distribution of more- and less-tacit patents within the United States. Specifically, as in Panel (C), we use the ratio of the coefficients of variation for more- to less-tacit patents, with higher values of this ratio indicating greater relative subnational spatial concentration for highly tacit patents. Our spatial units are 1990-vintage Commuting Zones, which use commuting flow to divide the land mass of the lower 48 states into more than 700 local labor markets (Tolbert and Sizer, 1996). Panel (D) reveals changes in how more- and less-tacit patents are distributed across the country. Between 1940 and 1980, a ratio near one indicates that both highly tacit and less tacit patented inventions emerged from a wide

[7] As seen in the SOM Section C, Panel (C), in absolute terms this pattern is general and secular across inventions, though it is considerably stronger among more tacit patents.

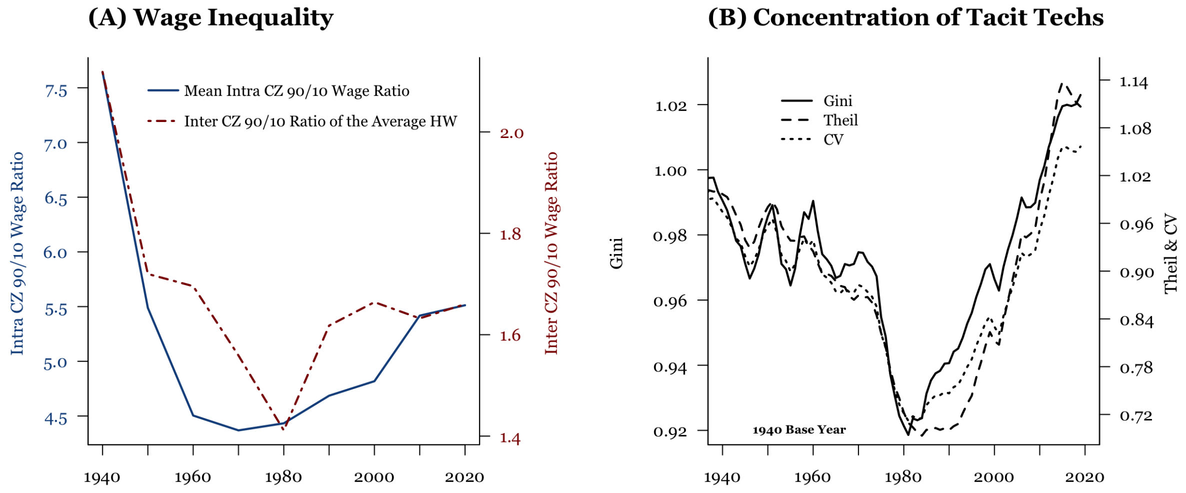
range of cities and regions. Then, after 1980, there is a clear upward trend in the ratio, indicating that highly tacit patents have come from a narrower range of places than in previous period.

We have now observed rising concentration after 1980 of the inventors, firms, technological domains and subnational locations from which tacit knowledge emanates. This provides a bridge to considering their possible economic effects. One such effect is inequality, since the emergence of new, major technologies is widely thought to reshape the distribution of income (Katz and Murphy, 1992; Autor et al., 2008a; Goldin and Katz, 2009). For instance, computers and other new technologies that began transforming the world of work sharply raised demand for workers performing particular kinds of nonroutine cognitive tasks (Autor et al., 2003). This demand grew at a rate that outstripped schools' capacities to produce new supplies of these workers (Goldin and Katz, 2009; Acemoglu and Restrepo, 2018). Excess demand raised wages for these workers, exacerbating interpersonal income inequality(Goldin et al., 2020), while the concentration of these highly-rewarded workers in certain regions has increased gaps in average incomes between cities and regions (Diamond, 2016). Over and above these well understood distributional effects, new and highly tacit technologies might be an additional source of income inequality for at least two reasons. First, when knowledge is tacit, education systems struggle to produce an increasing supply of knowledgeable workers, as there will be neither textbooks nor teachers for what has not yet been sufficiently codified. Second, technologies undergoing rushes of codification are likely to lie at the leading edge of 'general' shifts in the direction of technological knowledge. Individuals able to apply relevant (recently) tacit knowledge to these technologies ought to be able enjoy high relative wages, in the face of expanding demand for their scarce and hard-to-transmit know-how.

Patterns of U.S. interpersonal and interregional wage inequality are pictured in Figure 5, Panel (A), and are consistent with existing empirical work (Autor et al., 2008b; Gaubert et al., 2021; Kemeny and Storper, 2020). Using public microdata from a series of decadal and annual U.S. Census Bureau population surveys(Ruggles et al., 2022), this panel traces changes in the level of wage inequality between individuals in commuting zones, specifically the ratio of hours-adjusted annual income for an individual at the local 90th percentile to that of an individual at the 10th percentile. Inequality on this basis declined between around 1940 and 1970, briefly stabilized, and after 1980, grew for the next 40 years. Meanwhile, inter-regional inequality – the ratio of the income level of a place whose average wages lie at the 90th percentile of the distribution of locations to that of the 10th – follows a broadly similar pattern, declining over the 1940-1980 period, and reversing course thereafter. In short, after a "great levelling" in the middle of the 20th century (Lindert and Williamson, 2017), prosperity in the US, as measured by income, has been increasingly unequally distributed among individuals and households, and much of it concentrated in selected places.

Figure 5, Panel (B) captures the spatial dynamics of tacit technology production explored in Figure 4, this time presented in absolute terms. Panel (B) shows tacit technologies are spreading out over the geography of the United States between 1940 and 1980, at which point they reverse

Figure 5: Income Inequality and Spatial Patterns in Tacit Knowledge Production in the U.S.



Note: The solid line in Panel (A) presents patterns of inequality between individuals within local labor markets (1990-vintage Commuting Zones, or CZs) in the United States, measured in terms of individual annual hours-adjusted wage and salary income. For observed decades between 1940 and 2020, the line captures the ratio between an wages at the 90th and 10th percentiles of the local wage distribution. The dashed line captures income inequality between commuting zones, measured as the ratio of the income of a place with average income at the 90th percentile of the distribution of locations to that of the 10th. Underlying data are public microdata from a series of decadal and annual population surveys drawn from IPUMS (Ruggles et al., 2022). Panel (B) uses several conventional measures of inequality – Gini, Theil and Coefficient of Variation (CV) – to capture changes in absolute levels of the spatial concentration of highly tacit patents (those with tacitness scores above the median value). Spatial units are 1990-vintage Commuting Zones.

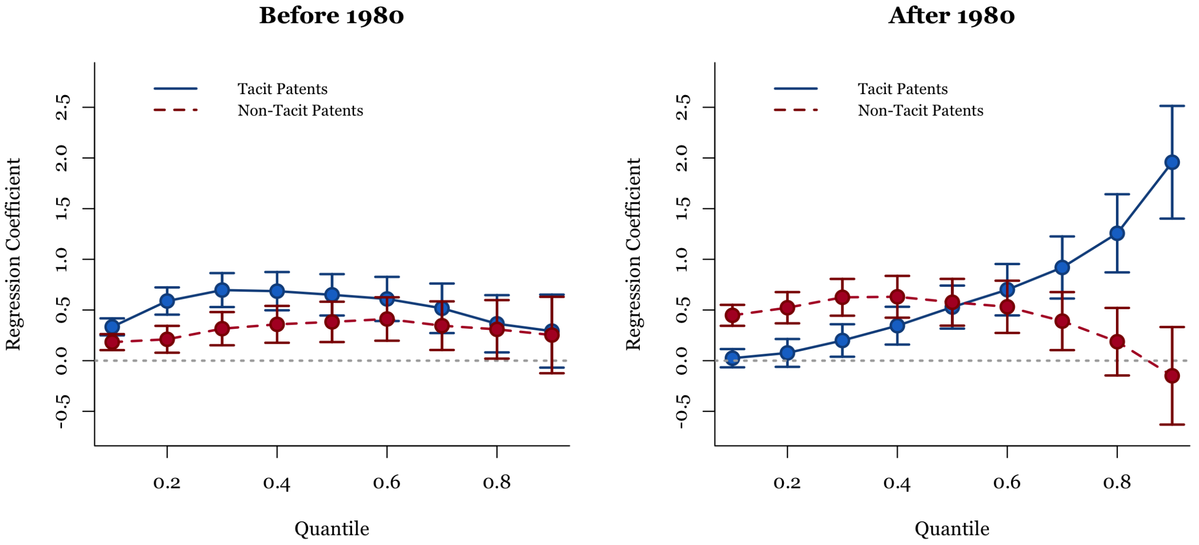
course, and become increasingly spatially concentrated, thus closely tracking the patterns of income inequality in Figure 5 Panel (A),

To investigate possible links between tacit patenting and income gaps between people and places we explore how local wage distributions change in response to the addition of more- and less-tacit patents. Given the sharp dividing line around 1980, we consider these relationships separately for 1940 to 1980, and then over the 1980-2020 period.[8]

The left panel of Figure 6 shows that, over the period 1940 to 1980, commuting zones that added more tacit patents experienced faster wage growth across all deciles of the local wage distribution. The one possible exception is for individuals at the 90th percentile of the distribution, for whom the 95% confidence interval crosses the zero line. Broadly though, in this earlier period, the economic rewards to highly-tacit patenting were widely spread across the wage distribution, in approximately equal measure. The same is true for less-highly tacit patents, though the average observed relationship is somewhat more modest for lower-income workers. Thus, over this early period, inventive localities – whether inventions were more or less tacit – enjoyed growth in average

[8] Detailed full-period results are in the SOM, Section D.1

Figure 6: Regression estimates of the relationship between local highly tacit patents and incomes at deciles of the local income distribution



Note: Point estimates and 95 percent confidence intervals are generated through a series of two-way fixed effects regression models, in which the dependent variable is local wages for a worker at a given quantile of the income distribution in a given commuting zone, estimated from Census population survey microdata, drawn from IPUMS (Ruggles et al., 2022). The key predictors are the number of tacit and non-tacit patents granted to inventors or assignees in that commuting zone in a given year. These variables are log transformed. Additional control variables include the log of local population size; the share of local workers between 16 and 65 who are foreign-born and who have not completed at least 4 years of college; the share of local workers who have completed at least 4 years of college; and a measure of local vulnerability to import competition and offshoring, estimated based on industry data, following the approach laid out in (Autor et al., 2013). Each regression includes fixed effects for year and commuting zone, with standard errors clustered by commuting zone.

wages, with such growth broadly shared among workers at all levels of income.

After 1980, as seen in the right panel of Figure 6, the patterns look different. Between 1980 and 2020, the addition of highly tacit patents offered no gains to workers at lower points in the local income distribution. Meanwhile, above the lowest quartile, the gains rise progressively with incomes. High-income workers are richly rewarded by the addition of new tacit patents. These results do not confirm that it is precisely those workers on the frontlines of new tacit knowledge who gain from new more-highly tacit technologies, though this is plausibly in part the case. But it does show that, after 1980, there are increasingly unequal gains from the production of new, highly tacit technological knowledge, disproportionately rewarding already high-income workers. This pattern is not at all evident prior to 1980. In addition, after 1980, the distribution of rewards to highly-tacit technologies is distinctly different from that of less-tacit technologies. Patenting of less tacit technologies is linked to larger rewards for local lower-income workers, whereas workers at and above the 80th percentile of the income distribution may receive no wage benefits from the local invention of less-tacit patents. Thus, whereas more- and less-tacit technologies seemed to have similar effects on wages for all parts of the labor force over the 1940-1980 period, after 1980 the positive association of patenting with wages is only in evidence for highly tacit technologies.

4 Discussion

Tacitness and its progressive codification have long been central to the study of science and technology (Saviotti, 1998; Howells, 1996), and to the study of knowledge more generally (Polanyi, 1966). This paper is the first to develop a method of approximating this dynamic in a systematic, large-scale way. By observing rushes of codification in patenting, we offer a glimpse into the moments at which highly tacit new technologies emerge into the scientific lexicon, their ideas becoming woven into collective knowledge through codification.

We linked measures of tacitness to new data on patenting inventors and organizations, shedding light on the system of production of tacit knowledge over time. Between 1940 and 1980, highly- and less-tacit inventions are roughly equally well distributed among inventors, firms and other organizations, as well as over technological domains and geographical regions. Then, from 1980 onward, inventors of highly tacit patents become increasingly rare, and compared to less-tacit new technologies, they are increasingly concentrated in a small number of technological domains and geographical locations.

The economic rewards to inventiveness also change after 1980. Before this time, when local workers and firms produced either more- or less-tacit inventions, economic benefits were widely shared. After 1980, while all types of new ideas continue to generate economic benefits for the region where they are generated, tacit knowledge generation spurs wage growth for high-wage workers, where less tacit innovations reward the middle of the distribution. Given that tacit technology generation is increasingly geographically concentrated since 1980, this makes tacit knowledge a likely contributor to increasing inequality between people and places.

Our discovery that the system for producing tacit knowledge changed between the 1940-80 period and the subsequent decades is consistent with hypothesized patterns of technological change over much of the past century. The electro-mechanical industrial revolution of the late 19th and early 20th century continued to generate new ideas and new applications over the 1940-1980 period, through an expansion of manufacturing activity that embodied these maturing technologies. The information technology revolution was embryonic in this period, but not yet technologically consolidated enough to have widespread, large scale transformative effects on the economy. By the late 1970s, however, the programmable integrated circuit made information technology the basis of a full-fledged industrial revolution that has unfolded over the past few decades, perhaps analogous to the early phase of the electro-mechanical revolution from about 1880 to 1930. Parallel processes were at work in biomedical research and other domains. Currently, a number of technologies that were incipient until quite recently, such as artificial intelligence and gene editing, may be on the cusp of having revolutionary large-scale effects. We will then be able to observe the scaled-up production systems for the tacit knowledge and rushes of codification that are fueling such advances. Our research opens up avenues for investigating these possible effects. They include the future of tacitness itself, which may be altered if AI can substitute for human inventiveness, an open question at this time. They also include the organizations, people, domains, and locations from which tacit knowledge is emerging and the emerging economic reward from it. These are four critical areas of investigation – and there may be many others – to pursue by extending the work reported here.

References

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: An inverted-u relationship. *The quarterly journal of economics*, 120(2):701–728.

Autor, D. (2014). Polanyi's paradox and the shape of employment growth. National Bureau of Economic Research Working Paper 20485.

Autor, D., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American economic review*, 103(6):2121–68.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008a). Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008b). Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P., Rigby, D. L., and Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature human behaviour*, 4(3):248–254.

Bell, A. G. (1876). Researches in telephony. *Proceedings of the American Academy of Arts and Sciences*, XII.

Bell, A. G. (1877). Improvement in telegraphy. USPTO Patent No.174,465.

Bell, A. G. (1878). Improvement in speaking-telephones. USPTO Patent No.201,488.

Berlin, L. (2005). *The man behind the microchip: Robert Noyce and the invention of Silicon Valley.* Oxford University Press.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., and Tahoun, A. (2021). The diffusion of disruptive technologies. National Bureau of Economic Research Working Paper 28999.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–44.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2012). Robust inference with multiway clustering. *Journal of Business & Economic Statistics.*

Cancho, R. F. I. and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Card, D. (2009). Immigration and inequality. *American Economic Review*, 99(2):1–21.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

David, P. A. (1990). The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review*, 80(2):355–361.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Dorn, D. (2009). *Essays on inequality, spatial interaction, and the demand for skills*. PhD thesis, University of St. Gallen.

Field, A. J. (2011). *A great leap forward: 1930s depression and US economic growth*. Yale University Press.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379):eaao0185.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., et al. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539.

Gaubert, C., Kline, P. M., Vergara, D., and Yagan, D. (2021). Trends in US spatial inequality: Concentrating affluence and a democratization of poverty. National Bureau of Economic Research Working Paper No.28385.

Goldin, C., Katz, L. F., et al. (2020). Extending the race between education and technology. In *AEA Papers and Proceedings*, volume 110, pages 347–51.

Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. harvard university press.

Gorman, M. E. and Carlson, W. B. (1990). Interpreting invention as a cognitive process: The case of alexander graham bell, thomas edison, and the telephone. *Science, Technology, & Human Values*, 15(2):131–164.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. National Bureau of Economic Research Working Paper 8498.

Howells, J. (1996). Tacit knowledge. *Technology analysis & strategic management*, 8(2):91–106.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics*, 107(1):35–78.

Kemeny, T., Petralia, S., and Storper, M. (2022). Disruptive innovation and spatial inequality. *Regional Studies*, pages 1–18.

Kemeny, T. and Storper, M. (2020). The fall and rise of interregional inequality: Explaining shifts from convergence to divergence. *Scienze Regionali*, 19(2):175–198.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. (2015). PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213.

Lam, A. (2000). Tacit knowledge, organizational learning and societal institutions: An integrated framework. *Organization studies*, 21(3):487–513.

Landes, D. S. (2003). *The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present*. Cambridge University Press.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Amy, Z. Y., and Fleming, L. (2014). Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). *Research Policy*, 43(6):941–955.

Lindert, P. H. and Williamson, J. G. (2017). *Unequal Gains: American Growth and Inequality since 1700*, volume 62. Princeton University Press.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. (2005). *Economic transformations: general purpose technologies and long-term economic growth*. OUP Oxford.

Maddison, A. (2007). *Contours of the world economy 1-2030 AD: Essays in macro-economic history*. Oxford University Press.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., and Guellec, D. (2008). The oecd regpat database: a presentation.

Mathews, J. A. (2012). Reforming the international patent system. *Review of International Political Economy*, 19(1):169–180.

Mokyr, J. (1992). *The lever of riches: Technological creativity and economic progress*. Oxford University Press.

Mokyr, J. (2009). *The enlightened economy: an economic history of Britain, 1700-1850*. Yale University Press New Haven, CT.

Moser, P. and Nicholas, T. (2004). Was electricity a general purpose technology? evidence from historical patent citations. *American Economic Review*, 94(2):388–394.

Nelson, R. R. and Winter, S. G. (2002). Evolutionary theorizing in economics. *Journal of economic perspectives*, 16(2):23–46.

Palombi, L. (2009). Beyond recombinant technology: synthetic biology and patentable subject matter. *The Journal of World Intellectual Property*, 12(5):371–401.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144.

Petralia, S., Balland, P.-A., and Rigby, D. (2016a). Histpat dataset. *URL: http://dx. doi. org/10.7910/DVN/BPC15W*.

Petralia, S., Balland, P.-A., and Rigby, D. L. (2016b). Unveiling the geography of historical patents in the united states from 1836 to 1975. *Scientific Data*, 3.

Polanyi, M. (1966). *The tacit dimension*. University of Chicago press.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102.

Ruggles, S., Flood, S., Foster, S., Pacas, J., Schouweiler, M., and Sobek, M. (2021). IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS.

Ruggles, S., Flood, S., Goeken, R., Schouweiler, M., and Sobek, M. (2022). Ipums usa: Version 12.0 [dataset]. Minneapolis, MN, https://doi.org/10.18128/D010.V12.0.

Saviotti, P. P. (1998). On the dynamics of appropriability, of tacit and of codified knowledge. *Research policy*, 26(7-8):843–856.

Saxenian, A. (1996). *Regional advantage: Culture and competition in silicon valley and route 128, with a new preface by the author.* Harvard University Press.

Shi, F., Foster, J. G., and Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73–85.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1):65–94.

Teece, D. J. (1998). Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets. *California management review*, 40(3):55–79.

Tolbert, C. M. and Sizer, M. (1996). US commuting zones and labor market areas: A 1990 update. United States Department of Agriculture, Staff report.

United States Patent and Trademark Office (2020). Performance and accountability report. USPTO.

Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., and Berton, L. (2019). A multi-centrality index for graph-based keyword extraction. *Information Processing & Management*, 56(6):102063.

Supplementary Material

A Creating a new dataset of inventor and assignee names from historical patents documents

Although the full text of each patent document is made freely available by the United States Patent and Trademark Office (USPTO), they are not always available in a research-ready format. Structured datasets have been developed over the past years to bridge this gap, most prominently the National Bureau of Economic Research (NBER) Patent Citation Data File (Hall et al., 2001). This database contains detailed information on almost 3 million U.S. patents granted between January 1963 and December 1999, including all citations made to these patents between 1975 and 1999 (over 16 million). Another widely used database containing information of US patents is the Patent Network Dataverse, which provides longitude and latitude coordinates of inventor addresses for patents granted by the USPTO from 1975 to 2010(Li et al., 2014). In a similar fashion, for patents filed to the European Patent Office (EPO) or to the World Intellectual Property Organization (WIPO) between 1978 and 2011, the REGPAT dataset of the Organisation for Economic Co-operation and Development (OECD) provides inventor locations (Maraut et al., 2008). Meanwhile, the USPTO now provides full records of all patents granted since 1975 – the year when the USPTO began to record patents electronically.

Since the main objective of this paper is to study the long term evolution of tacit knowledge, we developed a new dataset that describes inventor and assignee names for all granted patent documents prior to 1975, and that complements the preexisting HistPat dataset (Petralia et al., 2016a), a well-structured, comprehensive, and geo-referenced dataset of historical patents in the United States covering the years 1836 to 1975. HistPat contains county-level geographical information for nearly all patents granted over this period.

We make use of optically recognized patent documents made available by Reed Tech and Google[9] in addition to two sets of digital documents provided by the Smithsonian Library[10]: the 'Annual report of the Commissioner of Patents'; and the 'Index of patents issued from the United States Patent Office'. We proceed in two steps. First, using full patent descriptions, we follow the same general methodological procedure described for HistPat (Petralia et al., 2016b), calibrating the model to retrieve names of the inventors and assignees instead of geographical information. Second, we scrape names of inventors and assignees from the optically recognized, better structured documents provided by the Smithsonian Library and described above.

[9] https://bulkdata.uspto.gov

[10] https://library.si.edu

We combine this information into a new database containing 6,680,152 entries, providing information on the name of the assignee or inventor for 94.23% of all patents granted from 1836 to 1975. We also provide information about the city of the inventor/organization (as it appears in the text), in addition to U.S. state , or country if foreign. This geographical information complements already existing databases(Petralia et al., 2016a). We also include the year the patent was granted. Lastly, we provide a flag variable that identifies whether the exact name of the inventor or the assignee was found in more than one original source, as described before. For the period of interest in this particular study (1920-1975), we find information about an inventor or an assignee for 96.23% of all patents granted in that period. The database is publicly available at: https://doi.org/10.7910/DVN/FQWKGF.

B Defining and measuring tacitness

B.1 What is tacit technological knowledge?

Perhaps the best-known description of tacit knowledge comes from chemist and philosopher Michael Polanyi (Polanyi, 1966), who declared that knowledge is tacit when we “know more than we can tell.” We share with Polanyi the idea that tacitness indicates an inability for a human agent to meaningfully and completely share their knowledge with others. Polanyi was expansive, considering the intrinsic, individualized embodiment of all types of human experience. Our investigation of tacitness focuses more narrowly on scientific and technological fields of knowledge.

Considering tacit knowledge in the realms of science and technology demands modification of Polanyi’s view of individual subjectivity. In scientific and technological fields, the goal is codification – a societal process that transforms once-tacit knowledge into a form that can be widely understood and communicated. From this vantage point, technological knowledge is tacit not when a particular individual cannot transmit it, but rather when society has not yet established widely accepted codes that describe the knowledge. What distinguishes knowledge in this state is that it can only be shared when agents have direct interpersonal contact. Different from Polanyi’s conceptualization then, technological knowledge can remain tacit even if certain individuals have the ability to tell. The characteristic of tacitness is that the telling is likely to be complex, costly and non-routine, as in a long hallway conversation among researchers. Note that in that conversation, the participants may indeed extensively deploy codes, but the codes alone cannot tell the complete story; anecdotally, the interaction might proceed in a hallway, cafe or lab, but not over email. One important consequence is that those unable to directly interact with the holders of tacit knowledge and engage in this complex and costly process are excluded from knowing.

We can think of codification as the obverse of tacitness. Notice the distinction here between tacit and merely ‘complex.’ Even if billions of individuals cannot explain the complex knowledge behind gravitational mechanics, there exist agreed-upon linguistic and mathematical representations of it.

The availability of these codes permits suitably educated individuals to impersonally understand the laws of gravity, making it available to anyone who masters the codes, which can be learned through educational programs that are widely available. The deeper principle here is that technological knowledge is tacit when it is embodied in individuals, but not yet codified in such a manner as to permit impersonal transmission.

To illustrate the emergence of tacit knowledge and its eventual codification, consider Alexander Graham Bell's invention of the telephone. Bell aimed to create a viable multiple telegraph, an electrical device in which the same wire is used to transmit multiple text messages between two points. He did not initially set out to build a 'telephone' – a term not in wide use at the time. Yet, by combining a system of tuning forks developed by the German scientist Hermann von Helmholtz, with devices that grew out of Bell's speech-therapy work with deaf children, and in addition with discoveries he made converting sound waves into fluctuating current, Bell produced something new: his induction 'telephone' (Gorman and Carlson, 1990; Bell, 1876). While it made use of some preexisting technical ideas, his selection of which pieces of prior knowledge to use, and how to recombine them were unique and embodied – a function of his particular mind and lived experiences. Prior to its patenting, much of the invention was tacitly held, existing chiefly in Bell's mind. The only people who knew that this kind of assemblage might be possible – like his assistant Thomas Watson – had to get it directly from Bell. Indeed, it is telling that Thomas Edison, chasing the same prize at the same time, took a very different technical path, rooted in his own experiences and heuristics (Gorman and Carlson, 1990).

Until he patented it, Bell's particular assemblage of new and existing concepts was not yet part of collective scientific knowledge. Through the text and diagrams included in his initial patents (Bell, 1877, 1878), new concepts, such as those involved in rendering sound as fluctuating current, were traced with sufficient detail and clarity as to allow others to reproduce his device.[11] The codification process around key ideas did not end with Bell's initial patents. Those patents placed concepts in the public domain, and this spurred new inventions and patents (and litigation) by Edison, Elisha Gray and many others. We can imagine this as a rush of codification, in which a quantity of tacit knowledge around a set of technical concepts is rapidly transformed into a quasi-public good. Hence, at the scale of a concept at least, it makes sense not to think of knowledge as

[11] Note that in this process, Bell's novel, tacit concepts relied in large part on knowledge elements that were already codified. But their assemblage and new purposes were not codified. Bell's pathway is not unusual. New inventions often build on previously codified concepts, assembled in novel ways. Indeed, codification demands some degree of connection to existing ideas. To see why, imagine that human beings stumble upon a device created by an advanced alien civilization, whose workings depend entirely on principles unfamiliar to humanity. Such an invention will be incomprehensible, precisely because we would lack an ability to link its constituent ideas to principles with which we are already familiar. At the opposite extreme, consider a technology made possible only by concepts that have been previously codified, and that are used in expected ways. We would not consider this to be a (new) invention; at the present moment, it is redundant to the existing stock of collective knowledge. In short: for an invention to be new and comprehensible, it must be both connected in some way to codified knowledge, but not entirely so.

consisting of two static states – either tacit or codified – but rather as a threshold to be crossed upon initial formal or codified description of an idea, with subsequent unbounded space for further translation of tacit to codified knowledge, in the way that inventors today continue to repurpose and extend ideas codified in Bell's initial patents.

To begin generalizing from this example, we deploy four terms that describe the evolution of tacit technological knowledge, listed below from micro to macro:

- **Concept**: the smallest unit of technological knowledge, representing a single idea that can in principle be either tacit or codified to a greater or lesser degree. At its most micro level, a single word can represent a concept. In what follows, therefore, we use the terms 'concept' and 'word' interchangeably.
- **Invention**: a collection of concepts, linked together in such a way that creates a unique or novel device, method, composition or process. An invention, which commonly includes a mix of new and existing concepts, provides a new context in which new and existing concepts interact. An invention could in principle be a novel recombination of a set of entirely pre-existing concepts.
- **Technological Domain**: a set of inventions that are interrelated through underlying shared common principles or features.
- **Collective Knowledge**: the cumulative sum of codified technological knowledge available to humanity at a given moment in time.

B.2 Tacit technological knowledge from a network perspective

We model these four elements and their interaction using network analysis. Consider the available edifice of human collective knowledge as a network, in which each node represents a concept. An edge in this network links two nodes and represents a relationship between the concepts. A single invention will involve a series of edges between nodes, some of them new. On this basis, let the available collective knowledge (Ω) be:

$$\Omega_t = (W_t, E_t)$$

where $W_t = \{w_1, \ldots, w_{N_t}\}$ is the set of available concepts at time t, and $E_t = \{(w_i, w_j)\}$ the set of connections between concepts. In our network representation of the available collective knowledge, W_t and E_t collectively characterize the entire history of codified technological knowledge and interrelationships between concepts up to time t.

Of course, this network is not static. When a new invention is codified, it disturbs the network by adding elements to W_t or E_t, such that:

$$\Omega_{t+1} = f\{\Omega_t, I_t\}$$

where $I_t = (w_t, e_t)$ corresponds to an analogous network representation of the structure of concepts that is used in the process of codifying the new invention. The evolution of this process is also governed by f, which maps the transition between states of the network. Since the embedding of concepts underpinning the new invention adds new concepts and/or new relationships between concepts, it implies a transformation in the state of collective knowledge. Bell's invention of the telephone introduced novel, formerly tacit concepts that, at the time of patenting, were not yet part of the edifice of codified collective knowledge. Meanwhile, this invention also created new associations with already existing concepts. In both cases, by codifying new, formerly tacit knowledge, the invention of the telephone disturbs the preexisting network representation of collective knowledge.[12]

B.3 Patents as a paper trail of tacitness and codification

Our discussion thus far has suggested that patents, or at least Bell's seminal patents, signal the codification of once-tacit technological knowledge. Consider that, by definition, a tacit concept cannot be directly observed. At scale at least, we might only hope to observe tacit knowledge with codes of one kind or another, and yet, the lack of such codes is precisely what distinguishes tacitness. However, just as with Bell's initial patents, the moment a new concept emerges from the minds of a small number of individuals into wider collective knowledge, it signals that tacit knowledge linked to that concept has begun to be codified. The moment in which a patent is granted is thus not merely the provision of an intellectual property right, it also marks a moment of transformation in which a certain bundle of concepts shifts from tacit to codified technological knowledge. Note, however, that a new patent does not imply complete codification of the concept at hand, nor its use. Indeed, a patent does not necessarily resolve all the ways the concept can be turned into products and services, nor all the processes necessary to make devices and products with the concept.

Patents are not the only kind of document enabling the sharing of freshly-created codes. Nonetheless, given our focus on technological forms of knowledge and their economic effects, patent texts are a sensible place to look to capture codification in motion. In order to be granted a patent, inventors must transparently demonstrate the novelty and non-obviousness of the invention. Recent patents are thus trace elements of newly codified concepts and relationships; by implication, until the patent application is made, some concepts described in the application were held tacitly by a limited community of those in the know.

[12] In this regard, there is some similarity between our approach and conceptions of scientific knowledge as dynamic network in Shi et al.(Shi et al., 2015).

The patent system offers advantages in identifying the dynamic between tacit and codified technological knowledge. In patent documents, we find individual technical words and bigrams (pairs of consecutively-appearing words), which we take to represent concepts. Each patent document itemizes a particular invention, whose novelty has been expertly vetted. Meanwhile, patent offices like the USPTO expend great energy to organize patents thematically into classes; we use these classes as proxies for technological domains. Finally, we can consider the sum of all relationships between concepts in patents to provide a broad summary picture of the state of collective applied technological knowledge at a given point in time.

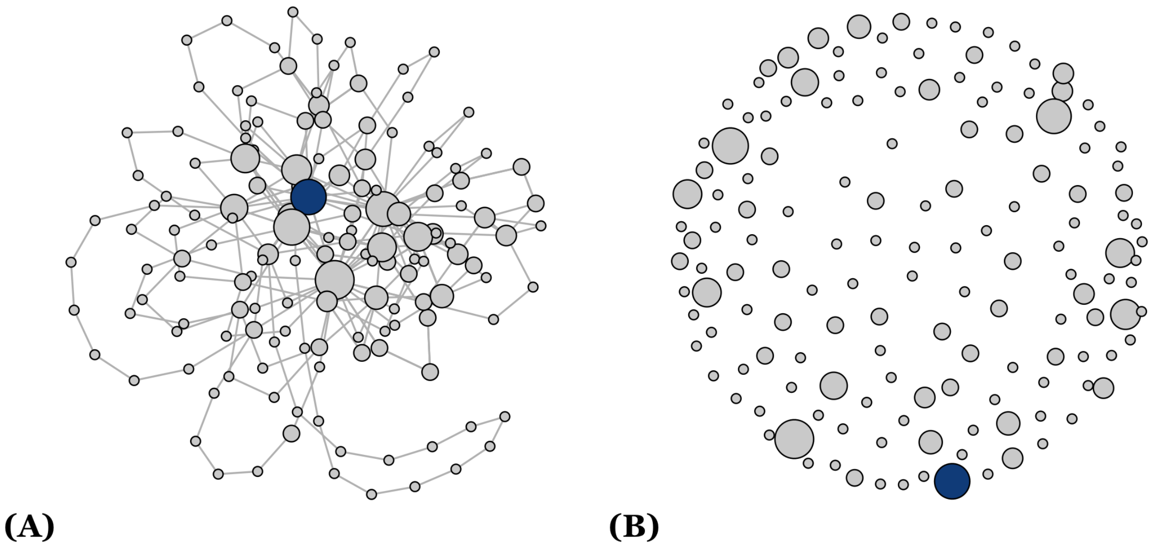
B.4 Representing meaning-making and codification in patents

A given concept can, in principle, contain an unbounded amount of codifiable, but as-yet tacit knowledge. To return to our initial example, Bell's patents started a process of codification around a set of concepts that enable the transformation, sending, and receiving of sound waves. This process did not end with Bell, but rather continues to this day, with new patents offering new applications, as well as novel combinations with other ideas. It follows that we want to be able to identify the moment that a particular concept is undergoing codification.

To respond to these challenges, we first need to think about how we know what a document is saying. For an individual patent document, consider how key concepts are contextualized in the text, where language confers meaning. Using the words found in Alexander Graham Bell's 1878 telephone patent, we define several possible approaches to this task in Figure SM.1. Each of them visualizes a particular way that words are linked to one another to confer meaning (the blue dot represents the word 'telephone'). Panel (A) is a network in which words in the patent are connected to each other by the order in which they were written. In this view, terms that come later in a sentence are used to explain and provide context to those found earlier. Panel (B) of Figure SM.1 starts from a different premise: that, lacking clear priors indicating the form of some universal semantics, all words in the document are interrelated. Each node is thereby connected to the other – though we do not show these lines, as they would overwhelm the figure. One could also motivate Panel (B) by assuming that we cannot identify which specific words are the main recipients of the context being provided in a given document's codification process, so we may as well consider all possibilities. Of course the assumptions underlying these representations are just two among many. In the case of Bell's initial patent document, we could imagine that almost all links should point toward the word *telephone*, on the basis that the concepts mobilized in the patent aim chiefly at describing the new invention. In addition, one might argue that paragraphs in a text consist of links between words (Cancho and Solé, 2001).

Notice that, even for an individual document, there is a potentially huge range of different ways that words may give meaning to concepts. Our research significantly magnifies this complexity, as we aim to capture codification not for a single patent document or concept, but for millions.

Figure SM.1: Semantic Network of Patent #201,488: A.G. Bell's Speaking Telephone



Notes: This graph provides two different network interpretations of Alexander Graham Bell's seminal telephone patent in 1878. Panel (A) displays a network representation in which words in the patent are connected to each other by the order in which they were written in the process of codifying the idea. Panel (B) considers that all words in the patent document are related to each other to some extent (we do not draw lines here, otherwise they would overtake the graph). The blue dot represents the word 'telephone'.

Our method for meeting this challenge is to use the frequency with which individual technical words and bigrams are used across patent documents. Consider a word that appears very infrequently across patent texts. With some confidence, we can conclude that the concept for which this word is a code is not well-integrated into the larger body of collective knowledge. It is a concept whose meaning is not highly codified. Now consider a technical word that is mentioned frequently across many patent documents. Such a concept ought to be relatively well-integrated into the edifice of collective knowledge. Moreover, when the number of appearances of such a word rise considerably over a relatively short period of time, the underlying concept is, in that period, experiencing a 'rush of codification.' This burst of mentions in patent texts signals that a concept's meaning is being rapidly stabilized, thus integrating it into the body of knowledge that can be communicated via impersonal channels.

Note that the heuristic we are proposing can be consistent with different theories of semantic meaning-making. Consider panel (A) of Figure SM.1, in this example one may argue that measures of network centrality may be a good indicator to identify concepts that are receiving meaning(Vega-Oliveros et al., 2019). After all, the word 'telephone' (blue dot) appears to be very central in this network. Note that a measure of network centrality of words such as the degree centrality will correspond one to one with the frequency with which words appear. This is because there is only one directional link for every concept in the document. The same argument can be made regarding the relationship between the frequency of appearance of concepts and the degree distribution of the network described in panel (B), provided that links are weighted accordingly. Therefore we consider that our heuristic may fit well across many possible meaning-making relationships between concepts, making frequency counts a reasonable basis from which to start measuring the codification of concepts.

B.5 Operationalization

Our primary source of information on collective technological knowledge comes from the detailed textual descriptions of every patent document granted by the USPTO since 1920. Despite the fact that inventions occur in many countries outside the U.S., because of its large market, patents granted in the U.S. represent an effective global gauge of invention.[13] In support of this idea, note that in 2020, more than half of USPTO patent applications were granted to non-Americans (United States Patent and Trademark Office, 2020). Usefully, the USPTO provides a taxonomy that allows to identify the technological domain of inventions, known as technological 'classes.'[14] Patent examiners assign each new patent to at least one class, according to the type of invention to which it claims rights. There are currently more than 600 different technological classes in use in the Cooperative Patent Classification (CPC) scheme. Whenever a new class is created, or an existing one redefined, all available patents are reclassified to maintain temporal consistency.

Patent documents are also useful because they contain detailed geographical information. We assign individual patents to locations using address information for inventors. For the 1920 to 1975 period, we rely on the HistPat dataset, which provides validated county-level information identifying the location of the inventor(s) and/or assignee(s) for 99.3% of all patents granted between 1836 and 1975 (Petralia et al., 2016b).[15] For the period from 1975 to 2020, we rely on similar data made available by the USPTO.[16]

A first step in creating our network of inter-related concepts is to identify a 'dictionary' that contains all technical words or phrases. We start from the collection of all individual words and bigrams in Wiktionary, an open-source and collaborative online dictionary that identifies words in all languages using definitions and descriptions in English.[17] Seeking to isolate technological concepts, we then search for the appearance of these words and bigrams in the universe of USPTO

[13] All patent documents granted since 1920 can be accessed at: https://bulkdata.uspto.gov

[14] http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products

[15] The latest version of HistPat can be downloaded at https://dataverse.harvard.edu/dataverse/HistPat. (Petralia et al., 2016b) contains a detailed documentation of the machine learning methodology used to create it and a set of tests to discard the existence of potential biases using manually collected data.

[16] Data is available at: https://www.patentsview.org/download/

[17] All their data is available to dowload at: https://dumps.wikimedia.org/enwiktionary/latest/.

patents granted between 1920 and 2020. We undertake several refinements to improve the usefulness of the dictionary. First, following common practice in textual analysis, we filter out 'stopwords' that add little substantive information, such as 'the', 'furthermore', or 'likewise.' We also remove words that appear either extremely infrequently (less than once per month), or too often (in more than 50% of all published patents in a given year). Further, because pre-1975 patents were digitized from paper documents using optical scanning techniques, in this earlier period we apply an additional cleaning routine to terms. To avoid including words that may look novel but that arise merely due to scanning errors, for the pre-1975 period we keep only words that also appear after 1975. Examples of these deleted words include 'examinei', which we presume is actually examiner, and 'terminai,' which should be terminal. Additionally, we excluded drug names and names of chemical compounds from our dictionary. We do so to avoid including words that may appear new and tacit, but that in fact represent merely new commercial names of existing drugs, or recombinations of already existing chemical recipes. Such relabelling is common practice in drug and chemical activities (Mathews, 2012; Palombi, 2009). To identify these terms, we rely on the PubChem repository, which provides information on chemical substances and their biological activities(Kim et al., 2015). Nevertheless, our results do not depend on excluding or including this set of words (see Section D.7) . After all of these cleaning routines, we obtained a final dictionary that covers the universe of USPTO patents between 1920 and 2020, containing 210,491 distinct technical words and bigrams.

To measure a concept's tacitness – or more precisely, the timing and extent with which it has experienced a burst of translation from tacit into codified knowledge – we consider two observable features of the distribution of words and bigrams in patent texts.

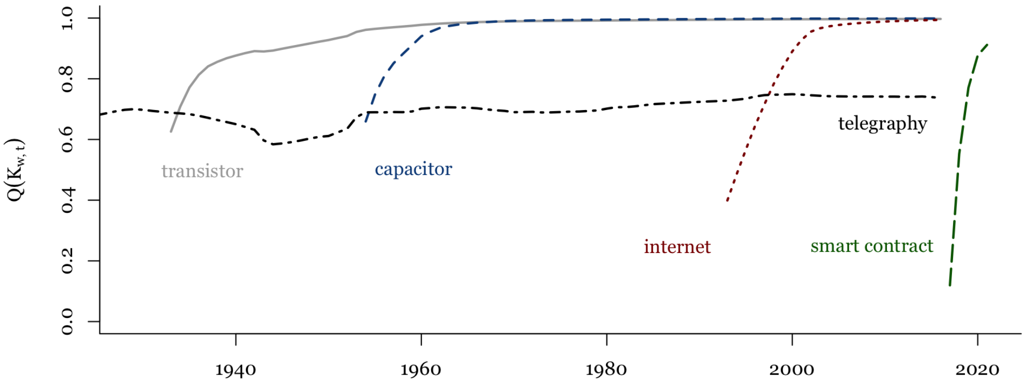
First, we consider that concepts that exhibit a rapid increase in the number of mentions across the universe of patent texts are likely to be undergoing a process of codification. We define K as the knowledge codified around each concept w up to year t as:

$$K_{w,t} = \sum_{i=1920}^{t} \sum_{d} O_{w,d,i}$$

where $O_{w,t}$ represents the number of occurrences of word (or bigram) w across patent documents d, from year 1920 up to year t. We then identify concepts that experience rapid growth in K based on the relative speed at which they accumulate knowledge around them, measured as $\Delta Q(K_{w,t})$, where $Q(.)$ represents the quantile each concept occupies in the distribution of mentions until year t and Δ represents its growth. Though all concepts are continuously transformed through the addition of new and reshaped relationships, some experience this transformation more rapidly.

For intuition on K, consider Figure SM.2, which illustrates how selected words evolve over the the study period, in terms of their movement along $Q(K_{w,t})$, plotted on the y-axis. This graph demonstrates that particular concepts follow distinctive pathways over time, in terms of the relative frequency with which they appear in patent documents. For instance, we can see that mentions of the word ‘transistor’ grow rapidly between 1920 and the first half of the 1930s, to later stabilize at a certain position in the distribution. Around 1970, while still commonly mentioned in patent documents, the term ’transistor’ wasn’t codified at a faster pace than the rest. Our interpretation of these patterns is that the concept we know today as ‘transistor’ underwent a process of rapid codification and re-definition in the 1920s and 1930s, as the term is being rapidly connected to many other new and existing concepts. Viewed another way, this is a period in which tacit knowledge linked to transistors is being rapidly generated but also rapidly transformed into codes. By 1940, transistors had become widely and well understood, even if they remain important in the larger process of technological change. In different periods, a similar time path is evident for concepts like ‘capacitor’, ‘internet’ and ‘smart contract’ (a concept frequently used in blockchain technologies), with the latter showing a much higher rate of codification than the rest. By contrast, by 1920 the word ‘telegraphy’ indicates an already a well-developed and stable meaning, which perhaps is not very different than the one we attribute it today.

Figure SM.2: The Codification of Concepts Over time



Notes: This graph shows the evolution over time of different words in the entire distribution of technological concepts. We show the percentile occupied by each word in each year, based on the value of $K_{w,t}$

A second observable feature of interest is the breadth with which concepts appear across patent texts. Our aim in measuring the breadth of mentions, B, is to distinguish redundant from non-redundant forms of codification, as certain patents – especially in pharmaceutical domains – repeatedly mention individual concepts, which suggest rushes of codification in terms of $\Delta Q(K_{w,t})$ that reflect only this narrow repetition. For example, in 2006 the word ‘dalbavancin’ was mentioned 1190 times, which represents the 0.56 percentile in mentions in that year, and an annual growth of 75%, and yet the many mentions of this word appear in only two patent documents. To account for this kind of issue, we define the breadth of mentions, B, as:

$$B_{w,t} = \sum_{d} max(I[O_{w,d,t} > 0])$$

where the indicator function $I[.]$ takes a value equal to one whenever a certain concept was used in a document. Therefore, $B_{w,t}$ captures the quantity of documents in which concept w appeared in year t. As above, we measure relative document occurrences using the quantile Q each concept occupies in a given year. 'Dalbavancin' would have a very low value of $Q(B_{w,t})$, reflecting the scant different contexts in which it appears.

Bringing these two measures together, we identify highly tacit (i.e. rapidly codifying) technical concepts in a moment in time as those whose words exhibit two features. First, highly tacit concepts are those that experience relatively rapid growth in the frequency distribution of all technical words and bigrams in patents. We define 'rapid growth' as at or above the 75th percentile in terms of growth in the frequency distribution of mentions ($Q(K_{w,t})$). Second, these appearances must be distributed across a relatively high number of patent documents, again defined using a threshold of at or above the 75th percentile. To take an example, in 2020, the word 'smart contracts' was at the 0.999 percentile of the rate of codification, and at the 0.865 percentile in the number of the number of patent documents where it was found. Hence we would define 'smart contracts' as a tacit concept in 2020.

Based on this procedure, Tables SM.1 & SM.2 describe the ten most highly tacit concepts for the years 1930 and 2020. For some contrast, we also list the ten most highly codified concepts, defined as those occupying the lowest ranks of $\Delta Q(K)$. In line with expectations, note that in Table SM.1, we can observe that highly tacit concepts are linked with electrical & electronic technologies, in line with the emergence of these new technologies (David, 1990). In Table SM.2, the most highly tacit concepts are vocabulary related to technologies featuring data mining, blockchain, and data science.

Table SM.1: Highly Codified and Tacit Concepts in 1930

	Highly Codified	Highly Tacit
1	shrub	television
2	studios	scanning
3	potassium sulfate	dispersions
4	cruising	polyhydric
5	grand total	rescue
6	clausen	photoelectric
7	strops	ethylene glycol
8	escape wheel	scan
9	hurd	short wave
10	appertain	time delay

Note: Highly codified concepts shown in this table are those with the highest rate of codification, provided that $Q(B_{w,t}) > 0.75$.

Table SM.2: Highly Codified and Tacit Concepts in 2020

	Highly Codified	Highly Tacit
1	quadrant plate	coreset
2	polyallomer	munyon
3	grov	kubernetes
4	grommeted	distributed ledgers
5	amonium	ethereum
6	aylor	smart contracts
7	methanols	autoencoder
8	soda fountain	numerologies
9	clothespins	immutably
10	tenite	entrepreneurship

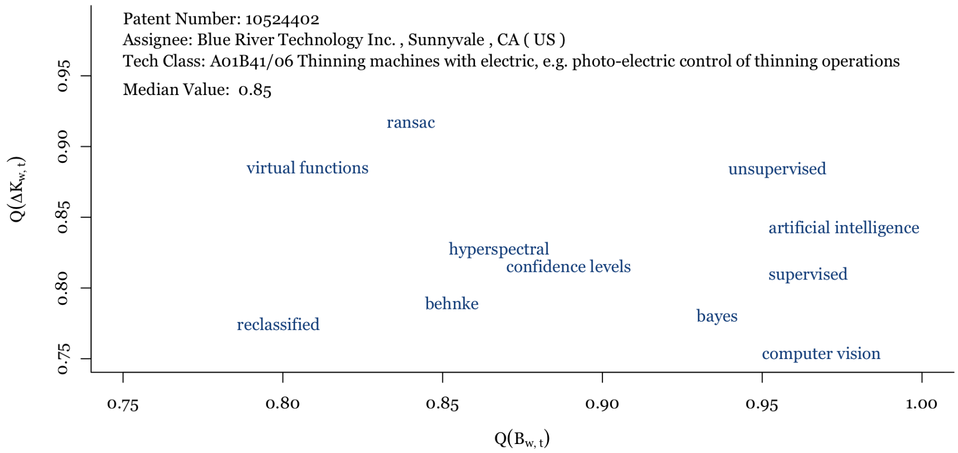
Note: Highly codified concepts shown in this table are those with the highest rate of codification, provided that $Q(B_{w,t}) > 0.75$.

Having identified more and less tacit concepts, we use them to create a distribution of patents in each year on the basis of their tacitness. For each patent we follow some simple procedures. We start by identifying the set of tacit concepts that appear in a patent (those at the top 25% in both measures). We then calculate the median of all values for those tacit concepts in a patent; we take the median of all pooled values because we cannot say, in principle, whether it is more important to have a greater rate of codification or to appear across a large number of documents. In practice, this means that a patent in which the median tacit concept is both experiencing a high relative increase in mentions, and these mentions appear across many patent documents will be ranked as more highly tacit than a patent in which the median concept is neither experiencing rapid growth in mentions nor appearing widely will be ranked as less tacit. Patents with no tacit words will be ranked least tacit. We limit our attention only to tacit concepts based on the idea that a patent containing highly tacit knowledge can have only a few tacit words surrounded by a larger number of words that are already relatively codified. Practically, in Bell's initial patent, the tacit concept of the 'telephone' must be defined in terms of ideas that are already relatively well-understood, or codified. Taking the median value across all words in a patent, tacit or not, produces a misclassification error, whereby patents like Bell's would in its day be incorrectly identified as being relatively well-codified.

Figure SM.3 shows an example of how we calculate the relative ranking of patent #10,524,402 – the very first patent granted in the first week of 2020. This patent incorporates machine learning techniques into agricultural machinery. The figure reveals a patent that contains tacit terms at various percentiles in the pace of codification and appearances across documents, with terms like 'unsupervised' as being rapidly codified in multiple contexts. In this example, the patent #10,524,402 receives a value of 0.85.

For the purposes of description, Tables SM.3 & SM.4 scale up further – to the level of technological domains (patent classes). We employ simple rules to determine tacitness ranking for patent

Figure SM.3: An example to illustrate how we measure the level of tacitness in a patent



Notes: This figure displays all tacit words found in patent #10,524,402. The horizontal axis shows the relative position of the words in terms of the amount of different contexts (documents) they were mentioned. The vertical axis shows the relative speed at which concepts were codified in 2020.

classes. In a given year, we deem individual patents highly tacit if they lie above the median level of tacitness across all patents.[18] Then we simply rank classes based on the percentage of their constituent patents are deemed highly tacit. Table SM.3 **??** shows that, in 1930, the top ten most tacit technological domains lie with the broad categories of electrical & electronic, chemical, and combustion engine technologies, in line with extensive documentation describing the emergence of these technologies(David, 1990; Lipsey et al., 2005; Bresnahan and Trajtenberg, 1995; Field, 2011; Moser and Nicholas, 2004). Meanwhile, the least tacit ten patent classes are largely mechanical in nature, including mature technologies whose potential for improvement is likely exhausted, such as "sewing" or "garment" related technologies.

Table SM.4, ranks tacit technological domains for the year 2020. The most highly tacit technologies in this last year of the study period lie mostly within the computer & communications category, in line with recent contributions around 'disruptive' innovation (Bloom et al., 2021), which identify the recent emergence of data science, artificial intelligence, and blockchain technologies. As in 1930, the least tacit technological classes appear to capture mature technologies, including "Rolling of Metal" and railway-related technologies. This descriptive evidence suggests a broad consistency between intuition and the results produced using our measure of tacitness. Further, class size doesn't seem to influence this ranking, as there are technological domains with many

[18] We get an almost identical ranking if we were to use the top 10% or top 25% as a cutoff for determining which patents are 'tacit'

Table SM.3: Most and least highly tacit technologies, 1930

CPC	Share	Total	Class Name
Code	Highly Tacit	Patents	
H03F	0.761	71	AMPLIFIERS
C08K	0.724	76	INORGANIC OR NON-MACROMOLECULAR ORGANIC SUBSTANCES AS COMPOUNDING INGREDIENTS
H04R	0.722	316	LOUDSPEAKERS, MICROPHONES, OR LIKE ACOUSTIC ELECTROMECHANICAL TRANSDUCERS
C10G	0.710	321	RECOVERY OF HYDROCARBON OILS FROM OIL-SHALE, OIL-SAND, OR GASES; REFINING MIXTURES
H04B	0.703	175	TRANSMISSION
F25B	0.663	184	REFRIGERATION, HEAT PUMP & COMBINED HEATING AND REFRIGERATION SYSTEMS
G06C	0.663	89	DIGITAL COMPUTERS W/MECHANICAL COMPUTATION
H02J	0.649	94	CIRCUIT ARRANGEMENTS FOR SUPPLYING, STORING OR DISTRIBUTING ELECTRIC POWER
H02H	0.638	94	EMERGENCY PROTECTIVE CIRCUIT ARRANGEMENTS
H04L	0.627	118	TELEGRAPHIC COMMUNICATION
...	...	...	...
A47H	0.086	70	FURNISHINGS FOR WINDOWS OR DOORS
A41F	0.122	98	GARMENT FASTENINGS; SUSPENDERS
A61F	0.130	115	FILTERS IMPLANTABLE INTO BLOOD VESSELS; PROSTHESES
D01H	0.133	98	SPINNING OR TWISTING
A41D	0.136	147	OUTERWEAR; PROTECTIVE GARMENTS; ACCESSORIES
A45C	0.143	126	PURSES; LUGGAGE; HAND CARRIED BAGS
A61G	0.145	55	TRANSPORT, PERSONAL CONVEYANCES FOR DISABLED PERSONS
B43L	0.150	60	ARTICLES FOR WRITING OR DRAWING UPON
B25D	0.151	93	PERCUSSIVE TOOLS
A01B	0.153	275	SOIL WORKING IN AGRICULTURE OR FORESTRY; AGRICULTURAL MACHINES OR IMPLEMENTS
A44C	0.158	101	PERSONAL ADORNMENTS, E.G. JEWELLERY; COINS

patents as well as few patents on each sides of the spectrum. Also, note that the appearance of technologies like 'HARVESTING; MOWING' in the top of 2020 is not a mistake; it identifies very new technologies arising in old established domains. The example discussed in Figure SM.3 is in this category as well.

Table SM.4: Most and least highly tacit technologies, 2020

CPC	Share	Total	Class Name
Code	Highly Tacit	Patents	
G06N	0.424	1,878	COMPUTING ARRANGEMENTS BASED ON SPECIFIC COMPUTATIONAL MODELS
G01W	0.343	67	METEOROLOGY
G05D	0.322	1,527	SYSTEMS FOR CONTROLLING OR REGULATING NON-ELECTRIC VARIABLES
H04L	0.307	21,307	TRANSMISSION OF DIGITAL INFORMATION, E.G. TELEGRAPHIC COMMUNICATION
A01D	0.304	418	HARVESTING; MOWING
G06Q	0.301	7,993	DATA PROCESSING SYSTEMS OR METHODS...
G10L	0.300	1,905	SPEECH ANALYSIS OR SYNTHESIS; SPEECH RECOGNITION; SPEECH OR VOICE PROCESSING..
G06F	0.291	36,067	ELECTRIC DIGITAL DATA PROCESSING
B60D	0.290	107	VEHICLE CONNECTIONS
G06T	0.286	6,090	IMAGE DATA PROCESSING OR GENERATION, IN GENERAL
...	...	...	...
A61P	0.010	193	SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS
G03G	0.012	1,904	ELECTROGRAPHY; ELECTROPHOTOGRAPHY; MAGNETOGRAPHY
A23C	0.012	81	DAIRY PRODUCTS, E.G. MILK, BUTTER OR CHEESE; MILK OR CHEESE SUBSTITUTES; MAKING THEREOF
D21F	0.019	52	PAPER-MAKING MACHINES; METHODS OF PRODUCING PAPER THEREON
A01H	0.024	1,047	NEW PLANTS OR ; NON-TRANSGENIC; PROCESSES FOR OBTAINING THEM...
C07J	0.027	74	STEROIDS
B21B	0.027	73	ROLLING OF METAL
D01F	0.032	94	CHEMICAL FEATURES IN THE MANUFACTURE OF ARTIFICIAL FILAMENTS, THREADS, FIBRES...
C07K	0.032	2,350	PEPTIDES
B61F	0.034	59	RAIL VEHICLE SUSPENSIONS
C07D	0.035	2,878	HETEROCYCLIC COMPOUNDS

C Concentration of tacit technological knowledge in people, organizations, domains and regions

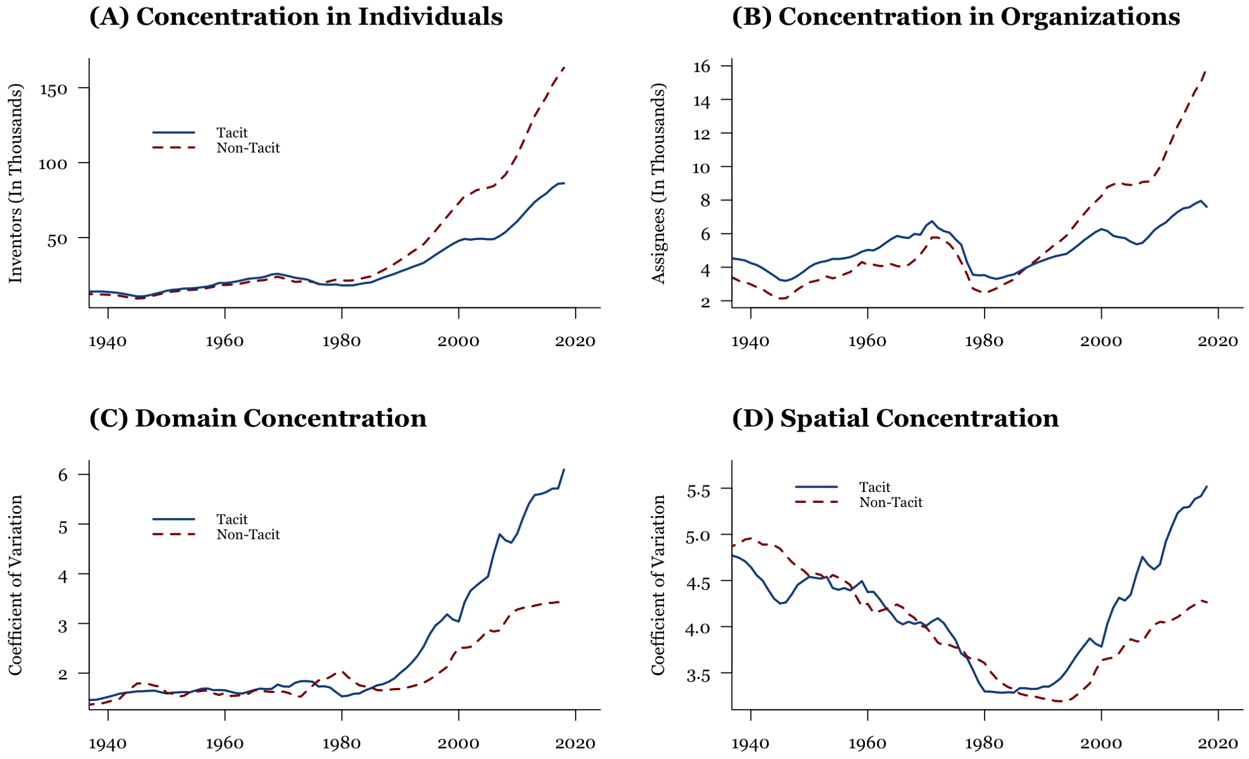
Knowledge is produced by human actors, who do so in a variety of settings: organizations, technological domains, and geographical regions, among others. In the main paper, we explore the changing distribution of highly tacit patents relative to less tacit patents in these different organizational, geographical and human units. In this section of the SOM, we explore the same information in greater detail, this time disaggregating to observe absolute levels separately for more and less tacit patents. This provides a somewhat different perspective from that obtained from Figure 4.

An historical shift around 1980 in several dimensions of the knowledge production settings is evident. In Panel (A) of this figure, we observe an upward trend in terms of concentration among inventors and assignees across both more- and less-tacit patents, though one that is considerably more pronounced among highly tacit patents. Considering organizations in Panel (B), concentration in less tacit patents remains roughly constant across the entire study period, whereas, after 1980, highly tacit patents become increasingly concentrated in absolute terms.

We observe steady levels of concentration in technological domains in Panel (C) until around 1980; thereafter both more- and less-tacit patents arise increasingly out of a more restricted set of patent classes, and this tendency is around twice as strong among highly tacit patents.

Finally, between 1940 and 1980 in Panel (D), we observe a spreading out of the locations from which emerge both highly- and less-tacit patents. Thereafter, both start to become more spatially concentrated, but highly tacit patents are becoming considerably more concentrated.

Figure SM.4: Tacit inventions in people, organizations, technological domains, and American regions, 1940–2020


(A) Concentration in Individuals

(B) Concentration in Organizations

(C) Domain Concentration

(D) Spatial Concentration

Note: Panels (A) and (B) track the number of inventors producing less and highly tacit USPTO patents, with the median tacitness rank dividing patents in a given year into two equal-sized groups. Panels (C) and (D) capture changes in the coefficients of variation for more- and less-highly tacit patents. Higher values of the CV indicate greater concentration in either Cooperative Patent Classification (CPC) classes (domains of technological knowledge) or 1990-vintage Commuting Zones (regions) for more highly tacit patents as compared with less tacit patents.

D Estimates of the relationship between tacit technological knowledge and the income distribution

It is widely agreed and documented that the advent of new technologies can affect the distribution of income in a society, insofar as it affects the relative demand for factors of production (labor, different types of labor, capital), as well as returns to ownership and property rights.

The standard framework in economics considers that the income distribution in an economy can change as the result of a race between new technologies, which alters the demand for workers, and the education system, which determines the supply of those workers (Autor et al., 2003; Goldin and Katz, 2009; Acemoglu and Restrepo, 2018). This standard framework is said to explain a considerable share of recent growth in income inequality between people (Goldin et al., 2020), as well as providing insights into the growing average income inequality between regional economies (Kemeny et al., 2022).

If highly tacit knowledge lies at the leading edge of technological changes, then holders of this knowledge may be very highly remunerated in relation to holders of knowledge that is more codified and more widely accessible. The nature of tacit knowledge adds some new dimensions to the standard framework of race between education and technology. First, education systems might be especially hindered in generating workers with the right tacit knowledge, since it is hard to educate for that which one cannot codify. Also, for some of the same reasons, firms seeking workers to apply the right tacit knowledge may find recruitment challenging, as activities involved in this work have not yet been fully defined into tasks, jobs and occupations. To take an example, at the dawn of the information technology revolution, degree programs in computer science were largely nonexistent, and the occupation 'computer programmer' was at best ill-defined. Leading firms sought workers using generalized aptitude tests delivered to degree holders across all disciplines, and even non-graduates (Berlin, 2005). Given these obstacles to increasing the effective supply of workers, as demand for the application of a technology rises, one should expect rising wages for workers that are perceived, or guessed, to have the right tacit abilities or knowledge.

To investigate possible links between tacit patenting and income gaps between people and places, we estimate how local wage distributions change in response to the addition of local patents that vary in terms of being more- or less-tacit. In light of the sharp divide observed in the distribution of tacit patents across people, domains, organizations and places before and after 1980, we consider the possibility that the relationships to economic effects may be specific to each 40-year period. Our baseline estimating equation is as follows:

$$y_{c,t} = \beta TacitPatents_{c,t} + \gamma OtherPatents_{c,t} + \theta X'_{it} + \phi_c + \nu_t + \mu_{c,t} \tag{1}$$

where y is income for location c in time t - either the average level for that location, or income at a specific percentile of the local income distribution. The key parameter to be estimated is β, which captures the association between highly tacit patents and the outcome. Meanwhile, we control for potential confounding from less tacit patents. The vector X' captures additional features of localities that prior work on income inequality suggest may shape the income distribution, and may be correlated with our key regressor. Specifically, in X' we include measures of population, to capture changes in local economies of scale (Combes et al., 2012). As a proxy for skilled worker supply, we measure the share of the local employment based consisting of workers who have obtained at least 4 years of college (Autor et al., 2008b). To capture potential effects from low-skilled immigration, we measure the share of workers born outside of the United States who have obtained less than four years of college (Card, 2009). Finally, using data from Autor and Dorn (Autor et al., 2013) we include a measure capturing the extent to which the local economy's industrial structure makes it vulnerable to offshoring to low-wage countries. Meanwhile, ϕ is a city-specific fixed effect that absorbs bias arising due to between-place differences in relatively stationary but unobserved characteristics of local economies, and ν captures the effects of unobserved time-specific shocks that exert uniform impacts across all locations, such as as business cycles. Finally, μ represents the standard disturbance term.

Both our dependent variable and control variables (with the exception of the offshorability measure) are built from public use extracts of population censuses, harmonized and made available to the public via IPUMS (Ruggles et al., 2021). These data are drawn from the largest available public use sample of the Decennial or American Community Survey in each available year. In practice this means a full count for 1940, five percent samples for 1960, 1980, 1990 and 2000, a three percent sample covering 2009-2011 (which for convenience we call 2010), and one-percent samples for 1970 and 2020. Adapting the probabilistic method described by Dorn (Dorn, 2009), we assign fractions of individuals in the Census to 1990-vintage commuting zones based on the proportion of each County Group, State Economic Area, or Public-Use Microdata Area that belongs in each commuting zone.

D.1 Main regression estimates

Table SM.5 summarizes our primary results, upon which Figure 6 in the main text is based. Results are estimated per Equation 1, using commuting zones' annual wage and salary income as the dependent variable and with a dummy variable coded 1 for patents granted after 1980, interacted with each kind of patent. Since in early years of our study period we can only track granting and not application dates, we use a five-year window forward from time t, such that we count any new patent granted in location c between years t and $t + 5$. Put simply, patent approval is not instantaneous; patents granted in $t + 5$ are likely to involve knowledge that is codifying well before the actual granting date, with attendant effects on work and its rewards. In Section D.6 we relax this assumption.

The regressand in Model (1) is a commuting zone's average wage and salary income, representing both how a worker with the local average income may be affected by tacit patenting, and at the same time capturing potential between-place inequality effects. This model shows that the average worker in the local wage distribution experiences wage growth in response to the addition of a highly tacit patent whose inventor(s) is located in that area. While this is true for tacit patents granted before and after 1980, the estimate of the relationship after 1980 is roughly twice the size of the obtained for the earlier period. The scale of returns to tacit patenting appear to be highly period specific. Meanwhile, less tacit patents also exert upward pressure on the average worker's wage, with coefficients that are similar across the two periods, and also, against a 95% confidence interval, statistically indistinguishable from pre-1980 tacit patents. Another reading of Model (1) is that, in the pre-1980 period, less- and more-tacit patents display a similar relationship

Table SM.5: Estimating the relationship between highly tacit patents and the income distribution in U.S. commuting zones

	Annual wage and salary income		
	(Mean)	(90th pct)	(10th pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.475***	0.292	0.334***
	(0.104)	(0.183)	(0.043)
Less tacit patents	0.389***	0.253	0.183***
	(0.100)	(0.192)	(0.040)
After 1980			
Highly Tacit Patents	0.838***	1.958***	0.024
	(0.130)	(0.283)	(0.046)
Less Tacit Patents	0.384***	−0.149	0.447***
	(0.126)	(0.245)	(0.053)
Population	0.524**	3.071***	−0.768***
	(0.258)	(0.508)	(0.088)
Share of Non-College Foreign Born	11.472***	33.054***	6.793***
	(2.188)	(4.361)	(0.839)
College Share	38.616***	57.491***	3.502***
	(2.420)	(5.561)	(0.684)
Offshorability	−1.108*	1.667	1.207***
	(0.662)	(1.287)	(0.251)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.953	0.941	0.885

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

to wages: innovation confers benefits for the average-wage worker and these benefits are unrelated to a patent's tacitness. After 1980, tacit patents yield distinctly larger rewards.

Models (2) and (3) investigate the idea that the effects of tacit patenting may differ for workers occupying different parts of the income distribution. One intuition is that highly-paid workers will include those workers most likely to be developing key tacit technologies, as well as those these inventions may complement, raising their productivity. In fact our results suggest a more nuanced picture. Considering workers at the 90th percentile of the local income distribution prior to 1980, the granting of new patents – whether more- or less-tacit – is not linked to wage growth. After 1980, tacit patenting is positively associated with wages; the coefficient on this estimate is large, suggesting a significant return to tacit patenting for these workers. By contrast, after 1980, less tacit patents are unrelated to the wages of workers at the 90th percentile.

Model (3) shifts attention to low-paid workers, specifically those at the 10th percentile of the income distribution. For this subset of the workforce, tacit patents granted before 1980 exhibit a positive association with wages, whereas after 1980, no such association is detected. For less tacit patents, we observe a positive relationship across the two periods. However, wage increases in response to less tacit patents for these low wage workers are considerably larger after 1980 than before.

Table SM.6 expands on these results, considering each individual decile of the local income distribution – in line with Figure 6. Across the income distribution, the addition of new local highly tacit patents is relatively consistently linked with increases in wages. However, the relationship is varies by period. Prior to 1980, new tacit patents are associated with higher wages for most workers, except for those at the 90th percentile, with coefficients highest for workers between the 20th and 70th percentiles. After 1980, the returns to tacit patenting rise with incomes, with no effects below the 20th percentile, and very large rewards for workers at or above the 80th. Less-tacit patents offer more consistent effects; larger after 1980 across the distribution, and in both periods weak or nonexistent for high wage workers.

Table SM.6: Estimating the relationship between highly tacit patents and the income distribution in U.S. commuting zones, by decile of the local income distribution.

	Annual wage and salary income								
	(10 pct)	(20 pct)	(30 pct)	(40 pct)	(50 pct)	(60 pct)	(70 pct)	(80 pct)	(90 pct)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Before 1980									
Highly tacit patents	0.334***	0.589***	0.697***	0.686***	0.650***	0.609***	0.517***	0.364**	0.292
	(0.043)	(0.068)	(0.085)	(0.096)	(0.104)	(0.111)	(0.124)	(0.144)	(0.183)
Less tacit patents	0.183***	0.211***	0.316***	0.358***	0.383***	0.411***	0.345***	0.309**	0.253
	(0.040)	(0.067)	(0.083)	(0.093)	(0.101)	(0.109)	(0.122)	(0.147)	(0.192)
Before 1980									
Highly tacit patents	0.024	0.076	0.199**	0.345***	0.529***	0.700***	0.919***	1.257***	1.958***
	(0.046)	(0.070)	(0.082)	(0.095)	(0.108)	(0.129)	(0.156)	(0.196)	(0.283)
Less tacit patents	0.447***	0.523***	0.625***	0.630***	0.576***	0.531***	0.390***	0.187	-0.149
	(0.053)	(0.078)	(0.092)	(0.105)	(0.117)	(0.131)	(0.146)	(0.170)	(0.245)
Period	1940-2020	1940-2020	1940-2020	1940-2020	1940-2020	1940-2020	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,523	6,523	6,523	6,523	6,523	6,523	6,523	6,523	6,523
Adjusted R^2	0.885	0.888	0.901	0.914	0.924	0.932	0.938	0.941	0.941

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation adjusted to 2019. Errors are always clustered at CZ according to (Cameron et al., 2012). All measures of patents are log-transformed counts, based on the year a patent was granted, as well as the subsequent five years. As in Table SM.5, controls included are log of local population, the share of non-college foreign born workers, the share of college educated workers, and a measure of local offshorability.

D.2 Alternative regression estimates: Hours-adjusted annual wages

In Table SM.7, we re-run the analyses that underlie Table SM.5, this time using measures of annual wages that have been adjusted for usual hours worked, as reported in the Decennial and ACS. Although magnitudes vary after accounting for differences in the number of usual work hours, the broad pattern is strongly consistent with the main results: we observe a strongly periodized result. For the average worker, wages grow in association with added tacit local patents in a manner that is larger than for less tacit patents after 1980; prior to 1980, both more- and less-tacit patents are linked to similar-sized wage increases. As in the case of annual wages, the relationship depends on where in the wage distribution we look, with high-income workers enjoying large benefits from tacit patenting after 1980, and little to no benefits from patents of any kind prior to 1980. Meanwhile, low-income workers benefit from less tacit patents in both periods, while highly tacit patents positively affect the wages of those in the lower ranks of the distribution only in the 1940-80 period.

Table SM.7: Replication of baseline results (Table SM.5) using hours-adjusted annual wage and salary income

	Hours-adjusted annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.258***	0.115	0.238***
	(0.047)	(0.100)	(0.028)
Less tacit patents	0.210***	0.170*	0.140***
	(0.046)	(0.103)	(0.027)
After 1980			
Highly tacit patents	0.318***	0.775***	0.016
	(0.054)	(0.127)	(0.026)
Less tacit patents	0.186***	−0.055	0.219***
	(0.054)	(0.113)	(0.030)
Population	0.325***	1.904***	−0.398***
	(0.114)	(0.243)	(0.061)
Share of Non-College Foreign Born	5.723***	18.102***	2.480***
	(0.930)	(1.914)	(0.556)
College Share	17.813***	29.835***	3.707***
	(0.973)	(2.459)	(0.455)
Offshorability	−0.552*	−1.666**	−0.007
	(0.305)	(0.752)	(0.170)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.955	0.928	0.913

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Hours-adjusted annual wages are in thousands, inflation adjusted to 2019. Errors are always clustered at CZ according to (Cameron et al., 2012). All measures of patents are log-transformed counts, based on the year a patent was granted, as well as the subsequent five years.

D.3 Alternative regression estimates: Varying patent cutoffs of $\Delta Q(K_{w,t})$ and $Q(B_{w,t})$

In this robustness check, we consider how a more stringent threshold for $\Delta Q(K_{w,t})$ and $Q(B_{w,t})$ – respectively the amount of knowledge codified around a concept, and breadth of mentions of a concept – may affect the shape of the association between tacit patents and wages. This implies a narrowing of the range of concepts that we consider to be tacit, hence alters which patents count as being highly tacit. In the main results, we consider concepts that are at or above the 75th percentile in both $\Delta Q(K_{w,t})$ and $Q(B_{w,t})$; here we raise that threshold to at or above the 90th percentile.

Table SM.8: Replication of Baseline Results (Table SM.5) with an Alternative Threshold for $K_{w,t}$ and $B_{w,t}$ (90%)

	Annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.055	-0.437^{**}	0.345^{***}
	(0.104)	(0.184)	(0.039)
Less tacit patents	0.762^{***}	0.870^{***}	0.215^{***}
	(0.107)	(0.196)	(0.043)
After 1980			
Highly tacit patents	0.618^{***}	1.796^{***}	-0.064^{*}
	(0.119)	(0.257)	(0.036)
Less tacit patents	0.608^{***}	0.154	0.515^{***}
	(0.112)	(0.212)	(0.042)
Population	0.572^{**}	3.170^{***}	-0.797^{***}
	(0.257)	(0.507)	(0.088)
Share of Non-College Foreign Born	11.053^{***}	31.303^{***}	7.009^{***}
	(2.171)	(4.233)	(0.827)
College Share	37.617^{***}	53.735^{***}	3.914^{***}
	(2.369)	(5.320)	(0.672)
Offshorability	-1.055	1.766	1.214^{***}
	(0.663)	(1.279)	(0.251)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.953	0.941	0.886

Notes: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

This greater stringency produces results that are broadly similar to the baseline in Table SM.5, but with somewhat larger effect sizes. Tacit patents are no longer significantly linked to wages for the average worker prior to 1980, whereas after 1980, coefficients are almost twice as large.

D.4 Alternative regression estimates: using a more restrictive definition of highly tacit patents

In this section we vary the threshold for defining a patent as highly tacit. Our main results proceed from the rule that a patent is deemed highly tacit in a given year if it has a tacitness score that is above the median. Here we raise this threshold to the 75th percentile, and then the 90th percentile. For the former, this means that only a quarter of patents in a given year will be deemed highly tacit, whereas for the latter, only 10 percent. These represent a significant increase in the threshold for what we identify as a highly tacit patent. It also makes the group of 'less tacit' patents more internally heterogeneous, combining patents that may involve no tacit concepts whatsoever and those that include considerably more than a typical amount of tacit ideas.

Table SM.9: Replication of Baseline Results (Table SM.5): restricting tacit patents to those at or above the top 75th percentile of tacitness.

	Annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.440***	0.166	0.321***
	(0.100)	(0.183)	(0.041)
Less tacit patents	0.458***	0.375**	0.208***
	(0.096)	(0.185)	(0.039)
After 1980			
Highly tacit patents	0.869***	1.932***	−0.003
	(0.132)	(0.296)	(0.044)
Less tacit patents	0.426***	0.019	0.461***
	(0.120)	(0.231)	(0.049)
Population	0.550**	3.170***	−0.767***
	(0.258)	(0.508)	(0.088)
Share of Non-College Foreign Born	11.181***	32.454***	6.836***
	(2.183)	(4.334)	(0.835)
College Share	38.406***	56.896***	3.680***
	(2.395)	(5.502)	(0.676)
Offshorability	−1.144*	1.660	1.208***
	(0.661)	(1.289)	(0.252)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.953	0.941	0.885

*p<0.1; **p<0.05; ***p<0.01

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

Nonetheless, in Table SM.9 and Table SM.10, results remain closely related to the baseline, indicating that our choice of cutoff is not fundamentally driving the overall relationships observed. The clear pre/post 1980 distinction remains evident, with outsize and inequality-inducing rewards to tacit patents after 1980, alongside more 'balanced' and even progressive returns to inventive activity prior to 1980. The fact that, with these more restrictive definitions of tacit patents, the estimated coefficients for tacit patenting rise (especially after 1980), can be taken to support the measurement approach we take in this study.

Table SM.10: Replication of Baseline Results (Table SM.5) Restricting tacit patents to those with tacitness scores at or above the top 90th percentile of tacitness.

	Annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.374***	−0.036	0.351***
	(0.105)	(0.192)	(0.040)
Less tacit patents	0.559***	0.571***	0.224***
	(0.094)	(0.180)	(0.039)
After 1980			
Highly tacit patents	0.967***	2.247***	0.011
	(0.135)	(0.309)	(0.042)
Less tacit patents	0.452***	0.047	0.445***
	(0.112)	(0.219)	(0.045)
Population	0.600**	3.283***	−0.774***
	(0.256)	(0.499)	(0.088)
Share of Non-College Foreign Born	10.977***	31.695***	6.887***
	(2.166)	(4.214)	(0.828)
College Share	38.112***	55.520***	3.841***
	(2.376)	(5.427)	(0.668)
Offshorability	−1.107*	1.606	1.242***
	(0.658)	(1.268)	(0.251)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.954	0.941	0.885

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

D.5 Alternative regression estimates: Using non-U.S. vocabulary

A further issue we take up is the risk that there is a correlation between specific subnational regional economies and particular kinds of technical vocabulary. If a particular location has its own lexicon around a set of technologies, and these technologies take off, then our determination of tacit concepts may be a product of this cultural specificity, rather than emerging more‘purely’ from universal or aspatial scientific and technological developments. To ensure that our determination of tacit concepts is fully exogenous from the subnational geography of inventions, we develop an alternative dictionary of technological words and bigrams that is restricted to patent texts in which all inventors and assignees are located outside of the United States. From these approximately 40% of USPTO patents over the study period, we build a unique dictionary, thereafter following the same steps described in detail in SOM Section B. The result is an alternative set of highly- and less-tacit patents that we can be certain is not driven by any such location-specific lexicons.

Table SM.11: Replication of Baseline Results (Table SM.5) with only vocabulary Created Outside the US

	Annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.492***	0.395**	0.341***
	(0.095)	(0.164)	(0.042)
Less tacit patents	0.376***	0.168	0.166***
	(0.100)	(0.190)	(0.041)
After 1980			
Highly tacit patents	0.888***	2.049***	0.039
	(0.137)	(0.308)	(0.048)
Less tacit patents	0.344***	−0.205	0.423***
	(0.128)	(0.263)	(0.051)
Population	0.505*	3.006***	−0.753***
	(0.260)	(0.512)	(0.088)
Share of Non-College Foreign Born	11.457***	33.001***	6.808***
	(2.174)	(4.313)	(0.842)
College Share	38.603***	57.341***	3.538***
	(2.411)	(5.543)	(0.685)
Offshorability	−1.127*	1.610	1.198***
	(0.666)	(1.297)	(0.251)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.953	0.941	0.885

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

Table SM.11 describes replication of our baseline results, substituting these alternative tacitness measures. Results are closely comparable to the baseline estimates shown in Table SM.5. As a consequence, we can be confident that any correlation between the emergence of words and their geography is not driving the results described in the paper.

D.6 Alternative regression estimates: Using Backward Looking Measures of Patenting

Table SM.12: Replication of Baseline Results (Table SM.5) Using Backward Looking Measures of Patenting

	Annual wage and salary income		
	(Mean)	(90 pct)	(10 pct)
	(1)	(2)	(3)
Before 1980			
Highly tacit patents	0.312***	0.139	0.139***
	(0.112)	(0.201)	(0.045)
Less tacit patents	0.256**	−0.006	0.279***
	(0.109)	(0.203)	(0.045)
After 1980			
Highly tacit patents	0.706***	1.437***	0.072
	(0.134)	(0.280)	(0.050)
Less tacit patents	0.172	−0.131	0.302***
	(0.128)	(0.233)	(0.050)
Population	0.816***	3.461***	−0.675***
	(0.260)	(0.515)	(0.088)
Share of Non-College Foreign Born	12.062***	34.192***	6.890***
	(2.235)	(4.468)	(0.852)
College Share	40.308***	60.320***	3.813***
	(2.424)	(5.641)	(0.694)
Offshorability	−0.737	2.336*	1.273***
	(0.686)	(1.326)	(0.260)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.952	0.940	0.883

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

D.7 Alternative regression estimates: including Commercial Drug Names

Table SM.13: Replication of Baseline Results (Table SM.5) including Drug Names

	Annual Wage (Mean)	*Annual Wage* (90 pct)	*Annual Wage* (10 pct)
	(1)	(2)	(3)
Before 1980			
Number of Tacit Patents (in logs)	0.410***	0.137	0.316***
	(0.103)	(0.186)	(0.042)
Number of Other Patents (in logs)	0.471***	0.450**	0.200***
	(0.102)	(0.194)	(0.041)
Before 1980			
Number of Tacit Patents (in logs)	0.827***	1.945***	0.012
	(0.129)	(0.284)	(0.047)
Number of Other Patents (in logs)	0.399***	−0.104	0.447***
	(0.123)	(0.240)	(0.051)
Population (in logs)	0.537**	3.101***	−0.765***
	(0.258)	(0.508)	(0.088)
Share of Non-College Foreign Born	11.383***	32.783***	6.821***
	(2.179)	(4.338)	(0.837)
College Share	38.339***	57.007***	3.449***
	(2.417)	(5.542)	(0.685)
Offshorability	−1.125*	1.666	1.202***
	(0.663)	(1.292)	(0.251)
Period	1940-2020	1940-2020	1940-2020
Year & CZ FE	Yes	Yes	Yes
Observations	6,523	6,523	6,523
Adjusted R^2	0.953	0.941	0.885

*p<0.1; **p<0.05; ***p<0.01

Notes: *p<0.1; **p<0.05; ***p<0.01. Unit of observation is 1990-vintage Commuting Zones. Annual wages are in thousands, inflation-adjusted to 2019 levels. All measures of patents and population are log-transformed counts. Errors are always clustered at CZ according to (Cameron et al., 2012).

References

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: An inverted-u relationship. *The quarterly journal of economics*, 120(2):701–728.

Autor, D. (2014). Polanyi's paradox and the shape of employment growth. National Bureau of

Economic Research Working Paper 20485.

Autor, D., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American economic review*, 103(6):2121–68.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008a). Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008b). Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P., Rigby, D. L., and Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature human behaviour*, 4(3):248–254.

Bell, A. G. (1876). Researches in telephony. *Proceedings of the American Academy of Arts and Sciences*, XII.

Bell, A. G. (1877). Improvement in telegraphy. USPTO Patent No.174,465.

Bell, A. G. (1878). Improvement in speaking-telephones. USPTO Patent No.201,488.

Berlin, L. (2005). *The man behind the microchip: Robert Noyce and the invention of Silicon Valley.* Oxford University Press.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., and Tahoun, A. (2021). The diffusion of disruptive technologies. National Bureau of Economic Research Working Paper 28999.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–44.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2012). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*.

Cancho, R. F. I. and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Card, D. (2009). Immigration and inequality. *American Economic Review*, 99(2):1–21.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

David, P. A. (1990). The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review*, 80(2):355–361.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Dorn, D. (2009). *Essays on inequality, spatial interaction, and the demand for skills*. PhD thesis, University of St. Gallen.

Field, A. J. (2011). *A great leap forward: 1930s depression and US economic growth*. Yale University Press.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379):eaao0185.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., et al. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539.

Gaubert, C., Kline, P. M., Vergara, D., and Yagan, D. (2021). Trends in US spatial inequality: Concentrating affluence and a democratization of poverty. National Bureau of Economic Research Working Paper No.28385.

Goldin, C., Katz, L. F., et al. (2020). Extending the race between education and technology. In *AEA Papers and Proceedings*, volume 110, pages 347–51.

Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. harvard university press.

Gorman, M. E. and Carlson, W. B. (1990). Interpreting invention as a cognitive process: The case of alexander graham bell, thomas edison, and the telephone. *Science, Technology, & Human Values*, 15(2):131–164.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. National Bureau of Economic Research Working Paper 8498.

Howells, J. (1996). Tacit knowledge. *Technology analysis & strategic management*, 8(2):91–106.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics*, 107(1):35–78.

Kemeny, T., Petralia, S., and Storper, M. (2022). Disruptive innovation and spatial inequality. *Regional Studies*, pages 1–18.

Kemeny, T. and Storper, M. (2020). The fall and rise of interregional inequality: Explaining shifts from convergence to divergence. *Scienze Regionali*, 19(2):175–198.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. (2015). PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213.

Lam, A. (2000). Tacit knowledge, organizational learning and societal institutions: An integrated framework. *Organization studies*, 21(3):487–513.

Landes, D. S. (2003). *The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present*. Cambridge University Press.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Amy, Z. Y., and Fleming, L. (2014). Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). *Research Policy*, 43(6):941–955.

Lindert, P. H. and Williamson, J. G. (2017). *Unequal Gains: American Growth and Inequality since 1700*, volume 62. Princeton University Press.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. (2005). *Economic transformations: general purpose technologies and long-term economic growth*. OUP Oxford.

Maddison, A. (2007). *Contours of the world economy 1-2030 AD: Essays in macro-economic history*. Oxford University Press.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., and Guellec, D. (2008). The oecd regpat database: a presentation.

Mathews, J. A. (2012). Reforming the international patent system. *Review of International Political Economy*, 19(1):169–180.

Mokyr, J. (1992). *The lever of riches: Technological creativity and economic progress*. Oxford University Press.

Mokyr, J. (2009). *The enlightened economy: an economic history of Britain, 1700-1850*. Yale University Press New Haven, CT.

Moser, P. and Nicholas, T. (2004). Was electricity a general purpose technology? evidence from historical patent citations. *American Economic Review*, 94(2):388–394.

Nelson, R. R. and Winter, S. G. (2002). Evolutionary theorizing in economics. *Journal of economic perspectives*, 16(2):23–46.

Palombi, L. (2009). Beyond recombinant technology: synthetic biology and patentable subject matter. *The Journal of World Intellectual Property*, 12(5):371–401.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144.

Petralia, S., Balland, P.-A., and Rigby, D. (2016a). Histpat dataset. *URL: http://dx. doi. org/10.7910/DVN/BPC15W.*

Petralia, S., Balland, P.-A., and Rigby, D. L. (2016b). Unveiling the geography of historical patents in the united states from 1836 to 1975. *Scientific Data*, 3.

Polanyi, M. (1966). *The tacit dimension*. University of Chicago press.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102.

Ruggles, S., Flood, S., Foster, S., Pacas, J., Schouweiler, M., and Sobek, M. (2021). IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS.

Ruggles, S., Flood, S., Goeken, R., Schouweiler, M., and Sobek, M. (2022). Ipums usa: Version 12.0 [dataset]. Minneapolis, MN, https://doi.org/10.18128/D010.V12.0.

Saviotti, P. P. (1998). On the dynamics of appropriability, of tacit and of codified knowledge. *Research policy*, 26(7-8):843–856.

Saxenian, A. (1996). *Regional advantage: Culture and competition in silicon valley and route 128, with a new preface by the author*. Harvard University Press.

Shi, F., Foster, J. G., and Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73–85.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1):65–94.

Teece, D. J. (1998). Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets. *California management review*, 40(3):55–79.

Tolbert, C. M. and Sizer, M. (1996). US commuting zones and labor market areas: A 1990 update. United States Department of Agriculture, Staff report.

United States Patent and Trademark Office (2020). Performance and accountability report. USPTO.

Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., and Berton, L. (2019). A multi-centrality index for graph-based keyword extraction. *Information Processing & Management*, 56(6):102063.