Appendix A – Description of Machine Learning Methodology

The dataset, or "corpus", is represented as a matrix composed of word frequencies for each article (row) and word (column). Frequencies can be simple term counts, but following O'Callaghan et al. (2015) we adopt a log-based term frequency-inverse document frequency (TF-IDF) representation, which helps to counter the influence of words that appear more frequently throughout the corpus. "Stop words" are entirely removed from the corpus. The term stop words is used to describe words which are most commonly used in a particular language (for example the conjunctions like 'and', 'if', or 'when', and prepositions like 'to', 'with' or 'in'). Such words are unhelpful in understanding the content of the corpus and are therefore ignored. Stop words were sourced from http://www.ranks.nl/stopwords. The corpus is then stemmed to ensure words with the same base are not counted separately. [1] We recommend that stemming and other text manipulation be undertaken with great care and only by those fluent in the language. For example, text analysis of Dutch patents is complicated by the prevalence of compound words. We further recommend that translation, where necessary, only be performed after the application of machine learning techniques where any single word mistranslation is likely to appear incongruous and therefore easy to detect.

Two commonly used approaches are Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Both attempt to model documents as combinations of topics, where topics are defined by the prevalence of particular words. LDA, introduced by Blei et al. (2003), assumes a generative process for documents and estimates the distributions. NMF uses a linear-algebra fitting technique. In preliminary investigations we found that topics suggested by LDA were more difficult to interpret. For example, when using 80 topics the 10 most prominent words appearing in the first topic of each model were:

NMF: steam, engine, rotary, rotatory, marine, navigable, condense, power, vapour,

[1] For example, 'cultivate' and 'cultivating' have the same base 'cultivat' but different stems, and would be observed as unique words without stemming.

part

LDA: igniting, wove, progress, circle, lantern, mache, decorating, multiplying, insects, check

The NMF topic suggests it has grouped patents for steam engines or other mechanical inventions. By contrast, the LDA topic does not seem to have derived a singular technology group, as it consists of a number of non-similar keywords. One reason for this relative interpretability of topics may be related to the specialist language necessary to describe inventions. Investigations by O'Callaghan et al. (2015) find that NMF can produce more coherent topics with associated generality and suggest that it may be more suitable for such non-mainstream domains. A second reason may be related to the relatively low average number (5.6) of non-trivial words used in the patent dataset titles. In experiments conducted over 'short text' datasets (with average word count ranging from 3.4 to 14.3) Chen et al. (2019) found that NMF was inclined to produce better topics than LDA. We also found that LDA produced more extreme topics: those that were dominant in either a very small or very large number of documents. Topics which were dominant in only a small number of documents proved particularly ambiguous and were not suggestive of a generalizable patent class.

To understand how the NMF approach works, suppose we have a corpus – a collection of patents in this instance – containing m patent titles, each composed of a set of n unique words. This corpus is represented by the matrix C, where $c_{i,j}$ represents, for each document i, the number of occurrences of word j. NMF attempts to factorize the matrix by approximating it as the product of two smaller non-negative matrices. This is represented as:

$$AT \approx C \tag{1}$$

where matrix T represents how often each word occurs within each topic. The weights in matrix A reveal the extent to which a patent relates to each topic. Word associations define their topics, which allows them to be interpreted by the investigator for further classification.

The number of topics is calibrated manually. When using topic scores to classify

patents, the number of topics influences where each patent is assigned.[2] Initially, we generated topics in multiples of 20, and manually examined the results. Fewer topics were associated with less consistent word associations, while additional topics alleviated this inconsistency. To find the appropriate balance between the number of topics and consistency of word associations, we rely on three separate measures: the Residual Sum of Squares (RSS); Entropy scores; and Coherence scores. These are displayed in Figure A1. Future investigators, when working with datasets of a significantly different size, should recreate this process to determine their optimal number of topics.

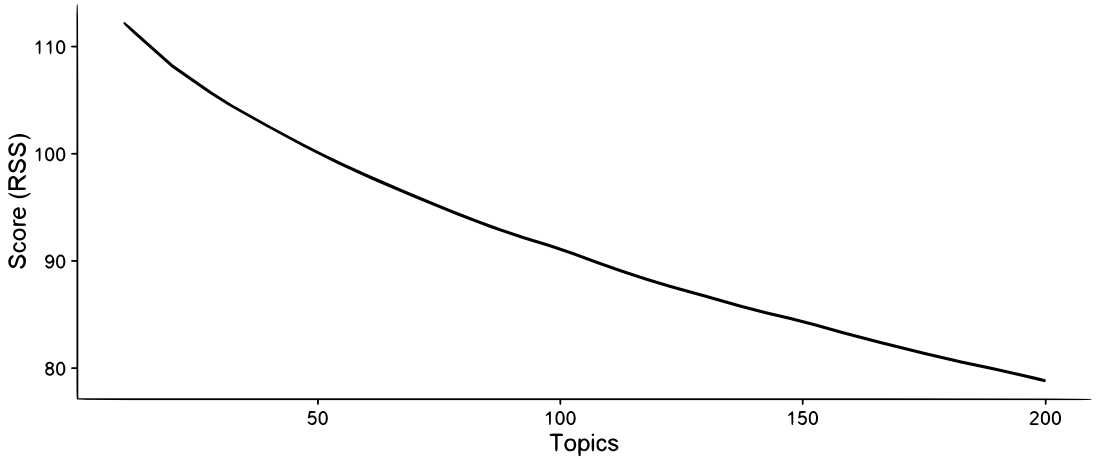
The RSS measures the quality of the approximation to the original document term frequency matrix, where a higher score suggests a less accurate representation. This metric decreases with each additional topic. In the case where there is a hidden number of groups, we may observe an improvement in the score once the number of topics reaches the number of these groups, with diminishing returns thereafter (Hutchins et al., 2008). Figure A1a shows the RSS scores to be decreasing in the number of topics, but at a marginal rate of decline. The slope of the curve becomes relatively flat between 50 and 150 topics, suggesting our optimal number of topics lies within this range.

Entropy is a measure of unpredictability. Information theory shows that changes in entropy proxy as a measure of information gain. Following Stevens et al. (2012), for topic model M partitioning data into t groups, where t is the number of topics, entropy can be measured as:

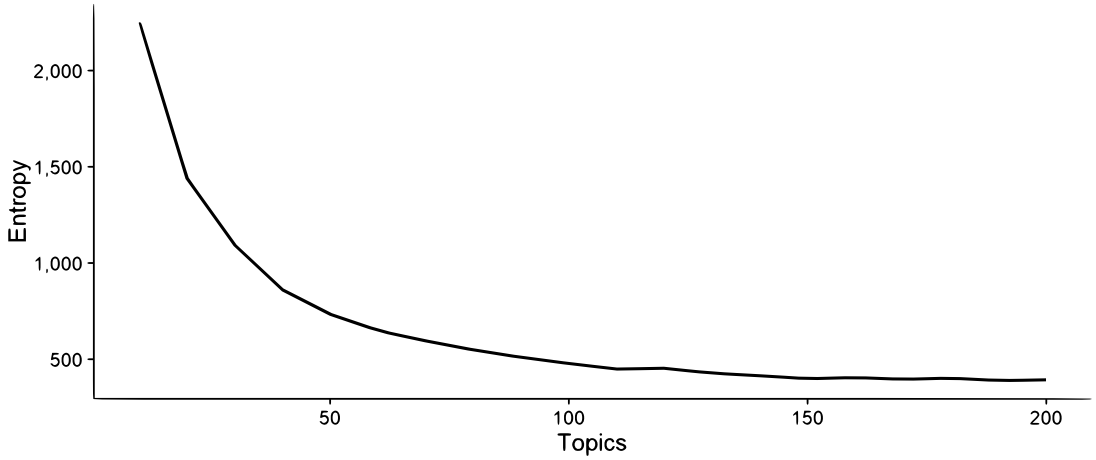
$$H(M) = \sum_{i=1}^{t} -P(i)logP(i) \tag{2}$$

Entropy therefore measures the amount of information gained from adding an additional topic. Figure A1b shows a negative association between the number of topics and information gain. A lower score suggests little information gain from adding an extra topic. The figure shows, for each additional topic, the new information received is diminishing. Between 10 and 60 topics is when the greatest information gain occurs. This steadily falls between 50 and 100, getting flatter as the number of topics passes 100. Information gain is relatively constant after 130 topics. Based on this measure, the

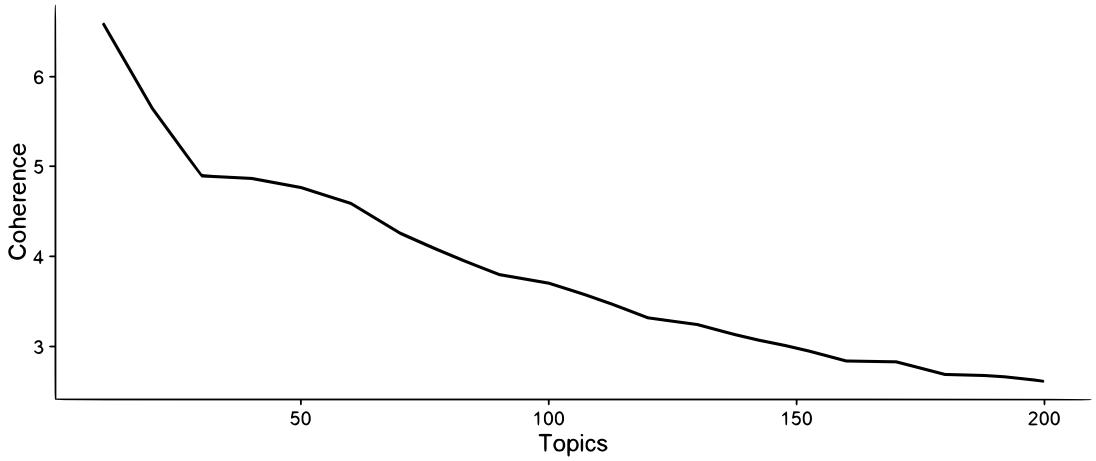
[2]We generate the optimal number of topics from our British dataset, described in Section **??**.



(a) *Residual Sum of Squares Scores Per Topic*



(b) *Entropy Scores Per Topic*



(c) *Coherence Scores Per Topic*

Figure A1: Measures for the Optimal Number of Topics

Source: Author's calculations using *A Cradle of Inventions: British Patents from 1617 to 1894* (2009)

optimal number of topics likely lies between 100 and 130, but closer to the upper bound.

Finally, we use Coherence-based scores. We can think of topics that make meaningful connections between words as being coherent. Measures of coherence are based on 'pairs of topic descriptor terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence' (O'Callaghan et al., 2015, p. 1). Stevens et al. (2012) consider measures of topic coherence which align with judgements by human investigators. One such measure is the "UMass" measure of Mimno et al. (2011). For topic T represented by the top n words t_i, the measure is defined as:

$$C(T) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} log \frac{D(t_i, t_j) + 1}{D(t_j)} \tag{3}$$

where $D(t_i)$ is the number of documents featuring word t_i, and $D(t_i, t_j)$ is the number of documents featuring both words t_i and t_j. For any given number of topics, we can then calculate the average topic coherence score.

Figure A1c displays the coherence scores. The overall trend suggests additional topics lead to less coherent associations. The figure shows a sharp decline in coherence between 10 and 30 topics. The scores steadily fall until 150 topics, where the slope becomes flatter. There is also a small increase in Coherence between 120 and 140 topics. This measure suggests the optimal number of topics falls between 120 and 150.

Based on the three metrics, we argue our optimal number of topics is 120. Each score indicates the range of 100-150 contains the optimal amount. We interpret the scores to point to 120 as an optimal number, however choosing 110 or 130 is unlikely to cause a substantial deviation in results. Thus, when we discuss or make use of topic analysis we will always use 120 topics. This does not mean we will have 120 distinct patent classes. On the contrary, topics are a means to derive common word associations that we will then classify according to our final patent schema.

Appendix B – Application of Taxonomy

To prepare the dataset for comparison, we assign each alternative taxonomy as described in the main paper. We assign our schema using our machine learning methodology. After deriving our 120 topics, we assigned one class per topic.[3] Our method creates each topic and assigns patents to them simultaneously. We then assign the topic's associated class. By assigning the top two topic scores to each patent, we can account for any potential overlap across technology groups. We denote these as "Topic-One" and "Topic-Two". We also manually classified the entire dataset, and compared our assignments with the machine's. Both authors did so independently. Either of our manually assigned classes matched either of the assigned topics in 93 per cent of cases.[4] The remaining seven per cent did not match because of too few unique words.

Table B1 presents a comparison of the schemas used in this study. Several classes appear across most of the taxonomies: Agriculture, Apparel, Chemicals, Engines/Power, Medical/Health, Metal, Military, Mining, Paper, and Textiles. For these commonly occurring classes, however, the number of assigned patents are not identical across taxonomies. The COI schema, for example, assigns 1,676 patents to Textiles, while our own Topic-One assigns 2,280; approximately 600 patents have been classified inconsistently across schemas. Food patents exhibit a similar inconsistency across existing schemas. COI lists 323 patents as Food, while NT lists 754 instead, and our Topic-One schema lists only 50. The majority of patents also receive a different Topic-Two assignment, suggesting the characteristics of many patented inventions overlap multiple technology groups. This suggests that patents should have more than one assigned classification.

We calculate Herfindahl-Hirschman (HHI) scores for each schema. HHI scores show how concentrated a particular taxonomy is. A higher score indicates a more skewed distribution of patents within a particular schema. For example, COI has the highest

[3]Where a topic was inconsistent in its word associations, it should be labelled 'Unclear' and then the patents assigned to it should be manually reviewed. In our analysis, no topics were labelled Unclear.

[4]We invited the senior patent examiner to manually classify a random sample of 250 patents according to our schema. We then checked their classification against the machine's and found a 70 per cent match.

Table B1: Comparison of Class Assignments

Cradle of Invention				Nuvolari				Topic-One				Topic-Two				Woodcroft			
Class	**Count**	**Percent**	**HHI**	**Class**	**Count**	**Percent**	**HHI**	**Class**	**Count**	**Percent**	**HHI**	**Class**	**Count**	**Percent**	**HHI**	**Class**	**Count**	**Percent**	**HHI**
AGR	442	3.21	0.001	Agriculture	455	3.30	0.001	Agriculture	597	4.33	0.002	Agriculture	493	3.58	0.001	Agriculture	483	3.55	0.001
BEV	278	2.02	0.000	Carriages	844	6.13	0.004	Apparel	105	0.76	0.000	Apparel	109	0.79	0.000	Apparel	179	1.31	0.000
CLO	279	2.02	0.000	Chemicals	1,152	8.36	0.007	Chemicals	1,189	8.63	0.007	Chemicals	990	7.19	0.005	Chemicals	151	1.11	0.000
COM	80	0.58	0.000	Clothing	344	2.50	0.001	Commodities	482	3.50	0.001	Commodities	217	1.57	0.000	Engines	1,018	7.47	0.006
DOM	1,642	11.92	0.014	Construction	641	4.65	0.002	Construction	564	4.09	0.002	Construction	778	5.65	0.003	Medical	237	1.74	0.000
FOO	323	2.34	0.001	Engines	1,714	12.44	0.015	Electricity	97	0.70	0.000	Electricity	60	0.44	0.000	Metal	432	3.17	0.001
IND	5,875	42.64	0.182	Food	754	5.47	0.003	Food	50	0.36	0.000	Food	77	0.56	0.000	Military	142	1.04	0.000
INS	458	3.32	0.001	Furniture	690	5.01	0.003	Hardware	1421	10.31	0.011	Hardware	1,689	12.26	0.015	Mining	40	0.29	0.000
MED	248	1.80	0.000	Glass	141	1.02	0.000	Health	85	0.62	0.000	Health	141	1.02	0.000	Paper	151	1.11	0.000
MIL	203	1.47	0.000	Hardware	879	6.38	0.004	Instruments	1153	8.37	0.007	Instruments	782	5.68	0.003	Textiles	1,323	9.71	0.009
MIN	207	1.50	0.000	Instruments	623	4.52	0.002	Machinery	666	4.83	0.002	Machinery	1,126	8.17	0.007				
MIS	15	0.11	0.000	Leather	224	1.63	0.000	Manufacture	459	3.33	0.001	Manufacture	1,075	7.80	0.006				
PAP	530	3.85	0.001	Manufacturing	736	5.34	0.003	Metal	517	3.75	0.001	Metal	484	3.51	0.001				
TEX	1,676	12.16	0.015	Medicines	287	2.08	0.000	Military	206	1.50	0.000	Military	116	0.84	0.000				
TRA	1,522	11.05	0.012	Metallurgy	719	5.22	0.003	Mining	166	1.20	0.000	Mining	253	1.84	0.000				
				Military	256	1.86	0.000	Paper	501	3.64	0.001	Paper	410	2.98	0.001				
				Mining	85	0.62	0.000	Power	1263	9.17	0.008	Power	1,464	10.63	0.011				
				Paper	504	3.66	0.001	Textiles	2,280	16.55	0.027	Textiles	1,863	13.52	0.018				
				Pottery	290	2.10	0.000	Transportation	1,080	7.84	0.006	Transportation	844	6.13	0.004				
				Ships	616	4.47	0.002	Utility	897	6.51	0.004	Utility	807	5.86	0.003				
				Textiles	1,824	13.24	0.018												
HHI			**0.229**				**0.070**				**0.083**				**0.080**				**0.026**

Notes: The table displays the Herfindahl-Hirschman Concentration ratios for each taxonomy. Count represents the total number of patents related to each class. This is then represented as a percentage. The individual class HHI scores are represented. The bottom row displays the HHI ratio for each taxonomy as a whole. For the 'Woodcroft' schema, we have included only those classes found in other schemas instead of the entire 146 classes. The HHI for Woodcroft is still calculated using the whole taxonomy.

Sources: Authors' calculations using data from *A Cradle of Inventions (2009)*, Nuvolari and Tartari (2011); Woodcroft (1860). All taxonomies cover 1700-1850.

associated HHI score at 0.229, while Woodcroft has the lowest at 0.026. Examining the COI schema shows 'Industry' accounts for 42 per cent of all British patents. No other schema has such a 'catch-all' class.

Appendix C – Patent Data

The patent data used in this study are the EPO's PATSTAT database – to help validate the construction of our patent taxonomy – and the entire population of British patents granted up to 1852 – to test for classification-specific econometric results. Designing and constructing any new time-invariant patent taxonomy requires robustness testing to ensure the taxonomy is capable of consistently classifying any and all historical patent data. PATSTAT is useful for robustness testing, because it contains the vast majority of all patents ever granted for Europe, the United States, and Japan. In addition, understanding whether the choice of patent taxonomy influences the results of any econometric analysis undertaken on patent data requires a dataset which has been classified according to multiple alternative taxonomies. The British patent data is useful for this comparative analysis, because it has been classified several times in the economic history literature.

Appendix C.1 – PATSTAT

PATSTAT is the EPO's comprehensive patent database. It is alleged to contain all patents ever granted, although its records are incomplete as shown (see main paper). Despite this, PATSTAT is still an incredibly useful dataset because it contains a wealth of bibliographic patent data. PATSTAT contains over 100 million patents from 90 different patenting authorities, with a large number of patents dating back to the nineteenth century. However, PATSTAT's historical coverage is much more complete from the early twentieth century.

The data contained in PATSTAT are digitised patent records, which can be found on Espacenet - an online collection of patent records from all member states of the EPO.

The patent records are recorded in their national language, and as such the digitised records maintain the original language. All patents held in PATSTAT contain their original patent title, and for the majority of patents the original specification is also digitised. PATSTAT also records the language patent titles are transcribed in, which is particularly useful for patents granted in countries which have multiple languages, such as Switzerland.

Because patents are recorded in their original language this provides a useful opportunity to extract a number of patent datasets covering a variety languages. As our goal is to construct a new patent taxonomy and methodology capable of classifying any and all patents according to a standard set of classes, having access to different national patent datasets allows us to test the versatility of our methods. If our method cannot classify patents in different languages in a consistent manner, then it cannot be considered a standardised or time-invariant taxonomy. For the remainder of this paper, we will be working with datasets from PATSTAT in their original language, and impose no translations during our methodology.

Appendix C.2 – British Patent Data

While PATSTAT is used for checking the versatility of our machine learning methodology, we use the historical British patent data to test whether the choice of classification influences the results of econometric analyses. This dataset contains all patents granted in Britain until the Patent Law Reform Act of 1852. Following the Reform Act, Bennet Woodcroft, then Superintendent of the Patent Office, meticulously collated all records for patents granted prior to the Act. In doing so, Woodcroft produced four tomes, each of which provides a wealth of information concerning all British patents granted. Of these, the tome 'Titles of Patents of Inventions', published in 1854, compiled the patent titles of all British patents, who they were granted to, the patentee's occupation(s), and their listed residence. Given this wealth of information, this tome has been digitised and compiled into a structured patent dataset on multiple occasions.

The Nuvolari and Tartari (2011) schema covers patents granted from 1617-1850.[5] According to their paper, the authors constructed their taxonomy based on a working paper version of Moser (2012), which relies on a taxonomy derived from the 1851 Crystal Palace Exhibition schema. The Crystal Palace schema comprises 30 technology groups, and was designed to encompass all possible inventions and submissions supplied to the 1851 Exhibition. Unfortunately, exactly how either of the aforementioned taxonomies were constructed is unclear, as is how any patent data have been classified.

A Cradle of Inventions: British Patents from 1617 to 1894 is a CD-ROM containing the entire population of British patents with the COI taxonomy assigned. Again, we were able to match up our schema to that from *A Cradle of Invention*. According to the insert provided with the CD, the taxonomy has been constructed as a simple means of aiding users to find relevant patent information. The insert states that the authors do not claim infallibility of the classification system, and that it is predicated on their interpretation of patent titles. In addition, class definitions are provided, which help us understand how patents may be classified, although the overall design of the taxonomy remains unclear.

Finally, Woodcroft (1860) is assigned based on a unique British patent identifier constructed by Woodcroft. When Woodcroft collated the British patent records, he assigned a unique number for each patent, for the purposes of linking patents through each of his four tomes. Each unique identifier is listed against at least one of the 246 classes in Woodcroft (1860), and so we were again able to match this data to ours. We can therefore be certain that each taxonomy has been accurately replicated. However, we are much less certain about how Woodcroft designed his taxonomy, but considering his role as Superintendent of the Patent Office, his goal may have been similar to those of patent examiners: group technologies based on functionality to help future inventors locate prior art. Woodcroft's schema is the largest by a considerable margin, which suggests the taxonomy was primarily for reference purposes.

[5]While their paper only covers patents granted until 1841, the dataset which they supplied had classified all patents up to 1850.

Appendix D – Additional Regressions

Here we present additional results for another four commonly examined patent characteristics from the economic history literature: patentee occupational status; the number of patents held by an inventor; the number of inventors listed per patent; and whether a patentee is considered an insider or an outsider.

Appendix D.1 – Patentee Occupational Status

To ascertain whether classification divergence is unique to examining the citations of patented inventions, we next examine patentee's occupations against patent classes. The innovation literature has examined the role of independent inventors and the types of industries they are likely to select into, or the types of inventions they are likely to produce (Schmookler, 1966; Khan and Sokoloff, 2004; Nicholas, 2010, 2011b; Khan, 2018). Our data allow us to conduct a similar examination. The patent data record the patentee's occupation alongside their name. This allows us to match occupations to a statistical measure of potential skills using the HISCLASS schema of Van Leeuwen and Maas (2011). This metric groups occupations based on their skills, whether they are manual or non-manual labour, and the degree of supervision required. For simplicity, we break the HISCLASS codes into manual versus non-manual occupations, following Klemp and Weisdorf (2012). Non-manual occupations are likely to be higher-skilled than their manual counterparts (Van Leeuwen and Maas, 2011).

We represent non-manual occupations using a dummy indicator variable. Consequently, a probit regression model is necessary to derive the probability of patent classes being associated with non-manual occupations. Our control variables constitute: whether the inventor had a prior patent; their nationality; and time controls. The explanatory variables are patent classes, with the baseline class being Agriculture.

Table D1 reports our results. Classification divergence still exists, however it is less severe for this particular characteristic. This may be due to the skewed distribution of non-manual occupations: approximately 75 per cent of occupations in the data are

classified as non-manual. Despite this, there exists variation because of class divergence.

Paper patents show a significant range in terms of coefficient size, for example. Under the COI schema, an average Paper patent is approximately 8.3 per cent more likely to be associated with a non-manual occupation, when compared to an Agriculture patent. The size of the result could be considered small, suggesting inventors of Paper patents were similarly skilled as inventors of Agriculture patents. However, the Woodcroft schema suggests non-manual occupations were, on average, 34 per cent more likely to produce Paper patents. While the conclusion remains similar, the contrast between coefficient sizes can lead to differing interpretations regarding the importance of human capital or skills when producing paper inventions.

Statistical significance also varies across taxonomies. The majority of patent classes report fluctuations between significance and non-significance. For example, Food patents are statistically significant at the one per cent level under the NT and COI schemas. The remaining schemas, however, are not statistically significant at conventional levels. Chemicals, Engines, Medicines, and Mining patents are the only ones to show no variation in statistical significance. This is in clear contrast to their variation observed against patent citations, suggesting divergences do not appear consistently when examining various patent characteristics.

The direction of association shows more stability compared with the previous set of results. Only Food, Instruments, and Paper patents show any variation in direction across taxonomies. The NT schema suggests Instruments patents were more likely to be associated with non-manual occupations compared to Agricultural patents. The remaining schemas, however, suggest the opposite: skilled individuals were less likely to produce instruments patents.

Appendix D.2 – Patent Stock

The third variable of interest is the stock of patents granted to patentees. This measures how many patents an individual has held in total each time they obtain a new patent. Patent stock is often used to control for other patent characteristics, and has

Table D1: Probit: Dependent Variable is a Dummy representing a Non-Manual Occupation

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	0.381***	0.189***	-	0.153***	0.099***	0.119***
	(0.065)	(0.045)	-	(0.046)	(0.023)	(0.015)
Clothing	0.061	0.051	0.011	0.102**	0.016	0.054*
	(0.046)	(0.040)	(0.041)	(0.040)	(0.035)	(0.031)
Engines	0.214***	0.183***	-	0.166***	0.081***	0.110***
	(0.037)	(0.040)	-	(0.043)	(0.028)	(0.019)
Food	-	0.118***	0.166***	-0.067	0.040	-0.015
	-	(0.042)	(0.050)	(0.103)	(0.052)	(0.032)
Instruments	-	0.044	-0.026	-0.017	-0.015	-0.036
	-	(0.055)	(0.059)	(0.058)	(0.041)	(0.026)
Medicines	0.271***	0.242***	0.237***	0.219***	0.133***	0.179***
	(0.040)	(0.036)	(0.043)	(0.038)	(0.030)	(0.027)
Metal	0.178***	0.164***	-	0.137***	0.069	0.086***
	(0.058)	(0.052)	-	(0.051)	(0.043)	(0.030)
Military	-0.064	-0.028	-0.048	-0.114*	-0.126**	-0.105***
	(0.051)	(0.064)	(0.064)	(0.067)	(0.053)	(0.034)
Mining	0.261***	0.210***	0.206***	0.148***	0.110***	0.104***
	(0.081)	(0.066)	(0.060)	(0.046)	(0.038)	(0.022)
Paper	0.341***	0.130***	0.083**	0.064	-0.028	0.016
	(0.052)	(0.043)	(0.040)	(0.046)	(0.026)	(0.012)
Textiles	-0.042	-0.033	-0.006	-0.050	-0.063	-0.046*
	(0.056)	(0.075)	(0.067)	(0.055)	(0.054)	(0.028)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	12,741	12,741	12,741	12,741	12,741	12,741
Pseudo R-Squared	0.132	0.0906	0.0728	0.0833	0.0698	0.0859

Notes: The table shows how the association between non-manual occupations and technology groups. The dependent variable is a dummy variable, where a value of 1 indicates a non-manual occupation. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as marginal effects at the means. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions (2009)* and Nuvolari and Tartari (2011); Woodcroft (1860). All datasets cover the period 1700-1850.

been studied in Dutton (1984); MacLeod (2002); Khan and Sokoff (2001); Khan (2015b). Patent stock is useful for observing the behaviour of professional inventors or patentees, who are likely to view patents more favourably or as a greater necessity than other inventors who do not exploit the patent system. These inventors can be more commonly thought of as "Economic men" (Dutton, 1984, p. 104-117), who respond to demand-side conditions. Observing what they patent can inform us about inventor perceptions of profitable avenues of invention.

Patent stock is represented by a simple count variable with a skewed distribution, as few individuals hold many patents while many hold few. Therefore, we use a negative binomial model, as when observing patent citations. The control variables constitute: inventor occupations; their nationality; and time controls. The explanatory variables are patent classes, with the baseline class being Agriculture. Table D2 reports our results.

Once again, we find evidence of classification divergence. Patent stock shows a relatively greater degree of divergence compared to our previous results. In terms of coefficient magnitude, there is substantial variation. Under the Woodcroft schema, inventors who held Mining patents, for example, are 60 per cent more likely to have had a greater stock of patents, suggesting inventors of Mining patents may have either deemed patents as necessary or earned enough profits from their patent stock to purchase additional patents. By contrast the COI schema suggests these inventors were only four per cent more likely to have held other patents, while the Topic-Two schema shows a magnitude less than one per cent.

Statistical significance also fluctuates across taxonomies. There is no single class to show consistently significant or non-significant results, which stands in contrast to the previous tables. For example, under the NT schema, Clothing patents are not statistically significantly different to Agriculture patents, while the Topic-One schema's result is significant at the ten per cent level, the COI's at the five per cent significance, and Woodcroft's at the one per cent significance. It would be difficult to conclude Clothing patents were different to Agricultural ones, given the results.

The direction of association also varies. Compared with prior results, fewer classes

Table D2: Negative Binomial: Dependent Variable is Patent Stock

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	-0.242	0.245	-	0.078	0.351***	-0.002
	(0.187)	(0.150)	-	(0.099)	(0.092)	(0.066)
Clothing	-0.429***	0.200	-0.302*	-0.369*	0.109	-0.399***
	(0.136)	(0.273)	(0.161)	(0.201)	(0.285)	(0.129)
Engines	0.085	0.244	-	0.188*	0.298**	0.072
	(0.175)	(0.191)	-	(0.096)	(0.126)	(0.073)
Food	-	0.233	-0.138	-0.029	-0.036	-0.265
	-	(0.166)	(0.128)	(0.247)	(0.206)	(0.182)
Instruments	-	0.216*	0.175	0.105	0.251**	-0.022
	-	(0.116)	(0.153)	(0.071)	(0.106)	(0.047)
Medicines	-0.476***	-0.282**	-0.495***	-0.381	0.004	-0.336
	(0.137)	(0.117)	(0.127)	(0.324)	(0.338)	(0.298)
Metallurgy	0.093	0.355**	-	0.205*	0.374***	0.149
	(0.187)	(0.166)	-	(0.118)	(0.107)	(0.118)
Military	0.151	0.455**	0.173	0.378***	0.290*	0.132
	(0.152)	(0.188)	(0.148)	(0.113)	(0.148)	(0.119)
Mining	0.606	0.367	0.075	0.238	-0.011	-0.176***
	(0.415)	(0.235)	(0.195)	(0.161)	(0.119)	(0.057)
Paper	-0.078	0.310*	0.244	0.206*	0.221*	0.032
	(0.302)	(0.164)	(0.162)	(0.115)	(0.114)	(0.091)
Textiles	0.404*	0.561**	0.347*	0.448**	0.475***	0.291**
	(0.234)	(0.218)	(0.205)	(0.189)	(0.128)	(0.146)
Constant	-0.480	-0.805*	-0.605	-0.650*	-0.777**	-0.590
	(0.430)	(0.451)	(0.389)	(0.363)	(0.374)	(0.369)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	13,347	13,347	13,347	13,347	13,347	13,347
R-squared	0.091	0.066	0.062	0.064	0.063	0.064

Notes: The table shows how the patent stock held by patentees at a given time varies by technology group. The dependent variable is the number of patents held by an inventor at the time of their latest patent grant. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as the difference in the logs of expected counts of the predictor variable. To translate this into a unit change, the coefficients need to be exponentiated. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions (2009)* and Nuvolari and Tartari (2011); Woodcroft (1860). All datasets cover 1700-1850.

are consistently positive or negative across schemas. For example, Military is consistently positive, while Clothing exhibits positive and negative associations. This suggests classification divergences are not consistent when examining different variables of interest.

Appendix D.3 – The Number of Inventors per Patent

The fourth characteristic we examine is the number of listed inventors per patent. Woodcroft included all named inventors when compiling patents granted prior to 1852. There are two plausible reasons why multiple inventors are listed: because either they helped develop the invention or they contributed to the cost of the patent. Inventor's had no claim over a patent right unless their name was also on that patent. Individuals who contributed to the development of an invention, either through their labour or their resources, would presumably have wished to retain a legal form of ownership over that invention. However, having an additional named inventor increased the cost of a patent by an undisclosed amount (Carpmael, 1842). Despite this, many patents have more than one named inventor. This suggests the additional cost was either outweighed by the benefit of retaining ownership or was sufficiently small so that inventors could effectively split the cost of the patent.

Obtaining a patent in England during our period of observation was expensive, with an average cost of approximately £100 in 1840 prices (Dutton, 1984). Given the high costs, either inventors had to be significantly wealthy or have access to additional funds, which could have been supplied by co-inventors or potential financiers. As part of the condition to extend financial resources to the prospective patentee, financiers may have requested their names be attached to the patent right. James Watt's famous patent, for example, was originally financed by his friends before Matthew Boulton became his financial partner (Bottomley, 2014a), but in Watt's case his friends were not listed on the patent.

The types of technologies which were likely to entail additional named inventors may reflect the larger fixed cost of making that invention (Jones, 2009) particularly where

technologies are far more complex and require a team of individuals to develop (Wuchty et al., 2007). Should additional named inventors be financiers, then correlations between technology groups and the number of named inventors may provide insights into the technologies perceived as profitable during our period of observation.

Table D3 reports our results, which show that classification divergence continues to exist. The divergence here is much milder compared to our prior results. Coefficient magnitude exhibits divergence. The largest fluctuations are observed in relation to Clothing, Medicines, Mining, and Textiles. For example, compared to Agricultural patents, Textiles patents are associated with about 12.5 per cent more named inventors under the Woodcroft schema. By contrast, the COI schema suggests that only seven per cent more named inventors are associated with Textile patents, while 'CombinedTopics' suggests only five per cent more. In several cases, coefficients are at least twice the size under one schema compared to another. Comparing two negative coefficients, for example, Clothing patents range from -0.077 up to -0.002, the latter being approximately 38 times larger in absolute terms.

Few classes exhibit any statistically significant coefficients across alternative taxonomies. The Medicines and Textiles classes exhibit the most consistent statistically significant coefficients, although evidence of divergence remains. Under the Woodcroft schema, Medicines patents are significant at the ten per cent level of significance, while both NT and Topic-One are significant at the one per cent level, and the Topic-Two and CombinedTopics schemas exhibit no statistical significance.

Direction of association also diverges considerably. For the majority of the common classes, the direction of association is consistently fluctuating. Only Instruments, Military, and Textiles patents report a consistently positive or negative correlation across taxonomies.

Appendix D.4 – Insiders versus Outsiders

Our next patent characteristic indicates whether the patentee is an 'insider': an inventor is an insider if their invention relates to their occupation. Physicians patenting medicinal

Table D3: Negative Binomial: The Dependent Variable is the Number of Inventors per Patent

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	0.059	-0.019	-	0.033	0.022	-0.010
	(0.082)	(0.044)	-	(0.034)	(0.022)	(0.022)
Clothing	0.055	-0.003	-0.044	-0.002	-0.043	-0.077***
	(0.056)	(0.041)	(0.043)	(0.049)	(0.031)	(0.027)
Engines	0.001	-0.005	-	0.008	0.017	-0.022*
	(0.039)	(0.041)	-	(0.032)	(0.021)	(0.013)
Food	-	-0.029	-0.037	-0.061	0.074	-0.026
	-	(0.039)	(0.039)	(0.058)	(0.079)	(0.053)
Instruments	-	-0.041	-0.033	-0.002	-0.014	-0.054***
	-	(0.040)	(0.041)	(0.034)	(0.020)	(0.013)
Medicines	-0.077*	-0.110***	-0.089**	-0.115***	0.018	-0.063
	(0.041)	(0.040)	(0.039)	(0.034)	(0.067)	(0.048)
Metal	0.049	0.011	-	0.070	0.073*	0.037
	(0.053)	(0.041)	-	(0.048)	(0.039)	(0.033)
Military	-0.087*	-0.095**	-0.080	-0.060*	-0.076*	-0.105***
	(0.052)	(0.043)	(0.049)	(0.035)	(0.044)	(0.023)
Mining	-0.083	0.048	-0.008	0.030	0.015	-0.014
	(0.069)	(0.080)	(0.054)	(0.047)	(0.036)	(0.022)
Paper	-0.018	-0.028	-0.005	0.012	-0.031	-0.052**
	(0.044)	(0.050)	(0.048)	(0.035)	(0.034)	(0.022)
Textiles	0.125***	0.112**	0.070	0.099**	0.095***	0.051**
	(0.041)	(0.046)	(0.045)	(0.044)	(0.029)	(0.025)
Constant	0.018	0.007	0.009	-0.015	-0.022	0.028
	(0.070)	(0.056)	(0.068)	(0.067)	(0.056)	(0.065)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	13,347	13,347	13,347	13,347	13,347	13,347
Pseudo R-Squared	0.00512	0.00347	0.00268	0.00290	0.00260	0.00299

Notes: The table shows whether the number of inventors listed per patent varies by technology group. The dependent variable is an ordinal variable indicating how many inventors were listed on a given patent. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as the difference in the logs of expected counts of the predictor variable. To translate this into a unit change, the coefficients need to be exponentiated. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions (2009)* and Nuvolari and Tartari (2011); Woodcroft (1860). All datasets cover 1700-1850.

inventions or engineers patenting engineering related inventions are considered insiders. Following the approach of Nuvolari and Tartari (2011), we construct a dummy variable to indicate whether a patentee's occupation matched the subject matter of their invention. In some cases, the occupation or subject matter is too vague to indicate whether the patentee was an insider. Because of this, we do not directly interpret a value of zero as reflecting an 'outsider'.

Whether an invention is developed by an insider has implications for how we understand the nature of invention. Jewkes et al. (1969) are amongst the earliest to argue that radical innovations are produced by outsiders, because such individuals are more willing to challenge accepted ideas. Insiders, by contrast, are too engrained into the technology to observe opportunities for radical advancement. O'Brien et al. (1996) have suggested that outsiders were responsible for significant advancements in textiles technology during the Industrial Revolution, while insiders were responsible for incremental improvements. Mokyr (2009) echoes this argument and considers outsiders responsible for "macro-inventions" – inventions responsible for opening up new technologies – and insiders responsible for "micro-inventions" – inventions which improved on existing technologies.

The results in Table D4 exhibit a relatively strong degree of divergence compared to the prior results. Coefficient magnitude displays considerable divergence. Clothing patents report the most extreme divergence. Under the Woodcroft schema, Clothing patents are 44-45 per cent more likely to be associated with an Insider than an Agricultural patent. But, the CombinedTopics schema suggests that Clothing patents are only eight per cent more likely. Our interpretation of the importance of insiders is then inconsistent; Woodcroft's schema suggests insiders are extremely important to Clothing innovation, while CombinedTopics suggests they are only marginally more important. Even in less extreme cases, such as Mining or Engine patents, there is significant divergence in terms of coefficient magnitude.

Statistical significance fluctuates considerably across taxonomies for several classes, most notably Food, Mining, and Paper patents. For those patent classes, significance

Table D4: Probit: The Dependent Variable is a Dummy representing an Insider

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	0.028	-0.039	-	-0.017	0.001	-0.112***
	(0.046)	(0.030)	-	(0.025)	(0.027)	(0.015)
Clothing	0.448***	0.228***	0.253***	0.237***	0.213***	0.082*
	(0.043)	(0.039)	(0.049)	(0.062)	(0.048)	(0.048)
Engines	-0.018	-0.002	-	0.044	-0.018	-0.080
	(0.030)	(0.034)	-	(0.094)	(0.097)	(0.051)
Food	-	-0.075***	-0.018	-0.071	-0.055	-0.124***
	-	(0.029)	(0.042)	(0.059)	(0.034)	(0.037)
Instruments	-	-0.035	0.012	0.041	0.007	-0.030
	-	(0.051)	(0.061)	(0.048)	(0.035)	(0.036)
Medicines	0.097	0.062	0.146**	0.139***	0.117***	0.057*
	(0.070)	(0.069)	(0.065)	(0.045)	(0.039)	(0.034)
Metal	0.133**	0.146***	-	0.139*	0.034	0.013
	(0.063)	(0.053)	-	(0.072)	(0.069)	(0.055)
Military	0.111*	0.070	0.101*	0.082**	-0.014	-0.013
	(0.057)	(0.061)	(0.059)	(0.039)	(0.039)	(0.030)
Mining	0.058	0.110*	0.141**	0.197***	0.145***	0.085***
	(0.084)	(0.056)	(0.060)	(0.038)	(0.031)	(0.027)
Paper	0.348***	0.177***	0.173***	0.035	0.029	-0.028
	(0.065)	(0.040)	(0.039)	(0.036)	(0.026)	(0.028)
Textiles	0.393***	0.318***	0.306***	0.271***	0.245***	0.174***
	(0.040)	(0.035)	(0.038)	(0.040)	(0.049)	(0.034)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	12,908	12,908	12,908	12,908	12,908	12,908
Pseudo R-Squared	0.225	0.196	0.180	0.182	0.179	0.191

Notes: The table shows whether a patent belonged to an insider as opposed to an outsider, and whether this varies by technology group. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as marginal effects at the means. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions (2009)* and Nuvolari and Tartari (2011); Woodcroft (1860). All datasets cover 1700-1850.

ranges from statistical significance at the one per cent level to no statistically significant association at all. This is most profound for Paper patents, where exactly half of the examined taxonomies report no significant association while the other half report an association at the one per cent level. By contrast, Clothing and Textile patents exhibit consistent levels of statistical significance.

The direction of association of coefficients is also subject to divergence, albeit to a much lesser degree. Military patents, for example, are inversely correlated with insiders under the Topic-Two and CombinedTopics schema, when compared to Agricultural patents. Conversely, the remaining taxonomies yield a positive correlation instead. Chemicals, Engines, and Instruments patents exhibit a similar division in terms of positive versus negative correlations.

Appendix E – Regression Plots for Common Patent Classes

The evidence presented in the main text supports our assertion that classification divergence exists, at the very least, in the innovation history literature. The results indicate that the choice of taxonomy can influence the size, significance, and direction of association of coefficients in a regression analysis of patent characteristics. However, this approach relies on contrasting the results of alternative patent schemas independently – no two schemas are directly examined in relation to each other.

To complement our results, this section directly contrasts patent classes from alternative schemas against each other. This approach allows us to understand whether the alternative schemas are correlated, which could highlight how similarly they classify patents. Observing correlations between taxonomies may be useful for predicting how much classification divergence we should expect in any regression analysis of patent characteristics. If correlations can reasonably predict how a particular taxonomy may yield diverging results compared to another, then this would be useful for identifying the amount of divergence in the literature. However, such an endeavour would require

access to other alternative schemas not discussed here, which is beyond the scope of our paper.

To identify the degree of correlation between taxonomies, we report regression plots which contrast the existing alternative patent schemas. Our model, described in equation 4, sets the Nuvolari-Tartari (NT) as our baseline schema which we then regress against each of the alternative schemas.[6] We run regressions for all six previously-presented patent characteristic metrics. Only results for patent quality and capital-saving metrics are reported; the remaining metrics show similar results and have been omitted for sake of brevity.

We present a series of plots of regression coefficients for each of the 12 common patent classes previously analysed. For a class to be considered 'common', it has to appear in at least three out of four of the comparable patent schemas. OLS regressions are used to estimate the degree of correlation between the alternative schemas for each common patent class and each patent characteristic. The regression equation is as follows:

$$NT_{ci} * Metric_i = \alpha_i + \beta S_{ci} * Metric_i + \mu_i \tag{4}$$

$Metric$ represents each of the six patent characteristics individually interacted with the NT common classes, denoted c, for each patent i that has an NT classification **??**. These characteristics are: the weighted number of citations per patent; the social class of an inventor's occupation; the inventor's current patent stock; the number of inventors listed per patent; whether an inventor is considered an 'insider'; and whether the patented invention is capital-saving.

The variable S denotes each of the other alternative schemas: COI, Woodcroft, Topic-One, Topic-Two, and CombinedTopics. Each alternative schema's classes, c, are also interacted with each of the six patent metrics separately, for all patents i which are classified according to those schemas. Each interacted alternative schema is then regressed against the interacted NT schema for each of the six characteristic metrics.

[6]We chose the NT schema to be our baseline since it contains all the common classes we wish to observe. We obtain similar results if we change the chosen baseline schema.

This approach allows us to observe the degree of correlation between patent schemas by focusing exclusively on whether each schema is capturing the same patents as the NT schema.

In each regression plot, a coefficient score of one suggests that classes from the NT schema and the comparison schema classify the same patents in the exact same way. By contrast, a score of zero suggests no correlation, which implies that neither schema is classifying the same patent into the same technology group. The confidence intervals for each patent schema's coefficient are also reported, which helps us to understand whether a correlation is spurious. We also include a reference line at a value of 0.5 for ease of interpretation.

It is important to note that the regression coefficients reported in Section 5 of the main text are for dummy variables; they are interpreted compared to the omitted category of Agricultural patents within each schema. Here the regression coefficients are interpreted against the baseline of the NT schema, rather than a single patent class within each schema. This creates difficulties in comparing the results from both methods, and we therefore make cautious comparisons between the main regression analyses and the regression plots.

Appendix E.1 – The Citations of Patented Inventions

In Table 8, classification divergence considerably influenced coefficient size, significance, and direction of association when observing the patent quality measure. Therefore, when examining patent citation measures, our interpretation of those measures is at risk of being influenced by the choice of schema.

To understand how correlated the alternative schemas are, and whether the degree of divergence is predictable, Figure E1 exhibits the regression plots from the OLS regression model for our patent quality measure: the weighted number of references per patent in the Woodcroft Reference Index. Each sub-figure represents the correlations between the alternative schema and the NT for each of the common classes. In each sub-figure, regressions are run separately and then reported collectively on the same plot.

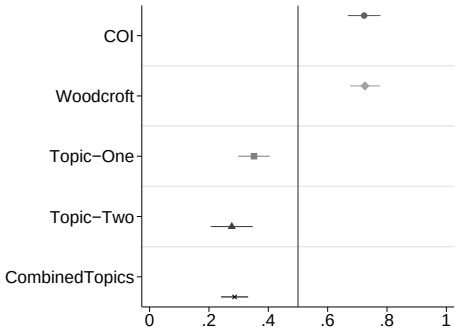
Observing sub-figure E1a, for example, shows both the COI and Woodcroft schemas are strongly correlated with the NT schema. By contrast, the Topic-One, Topic-Two, and CombinedTopics schemas are much less correlated with NT. To understand whether regression plots can be useful for predicting possible divergence, we contrast the plots for another patent class where the COI or Woodcroft schemas are also strongly correlated with NT. Medicine patents, for example, show a similar degree of correlation between our alternative schemas and NT compared to Agricultural patents: COI and Woodcroft are strongly correlated with NT in both instances. Therefore, we would expect very similar results for NT, COI, and Woodcroft in a regression analysis on patent quality. Referring to the results in Table 8, we do not observe such similarity. The Woodcroft coefficient is larger than the NT coefficient, while the COI coefficient is instead smaller. Furthermore, the Topic-One schema's coefficient is more similar in size to the NT coefficient, despite the lack of correlation observed in the regression plots.

For the majority sub-figures, the COI and Woodcroft schemas are strongly correlated with the NT schema, while the Topic-One, Topic-Two, and CombinedTopics schemas generally are not. This difference in correlations does not appear to predict classification divergence in our main results. For Chemicals, Metallurgy, and Textiles patents, the Topic-Two schema is not correlated with the NT schema compared to Woodcroft and COI, but Topic-Two coefficients in Table 8 are similar to the NT coefficients. This suggests that the regression plot measures are not reliable predictors for degree of classification divergence.

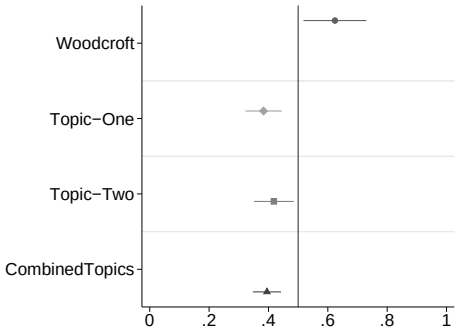
Appendix E.2 – Capital Saving Patents

The second metric we examine indicates whether a patent is intended to save on capital. The absolute divergence results for this metric are presented in Table 9. The degree of divergence for capital-saving patents is much more extreme compared to patent quality.

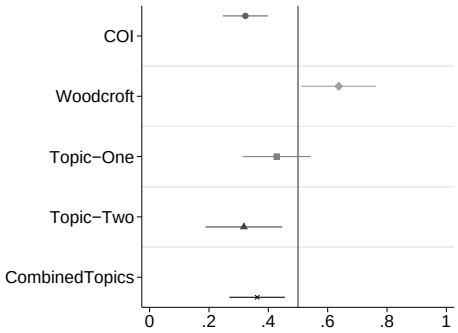
Figure E2 reports the regression plots for the capital-saving variable. Compared to the patent quality regression plots, there is a stark difference when observing capital-saving patents. While the general trend remains similar to the patent quality



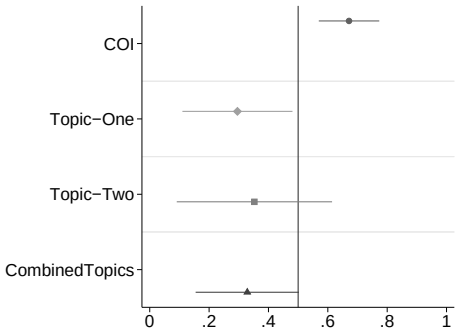
(a) Agricultural patents



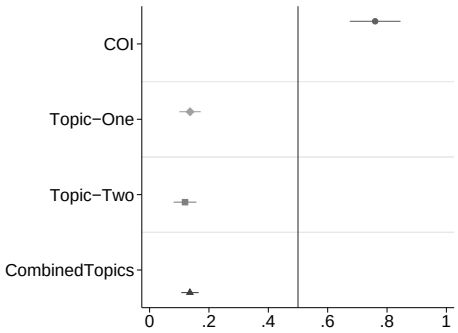
(b) Chemical patents



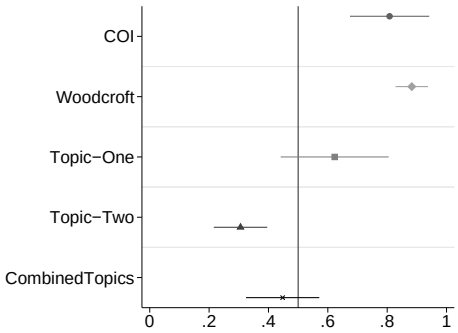
(c) Clothing patents



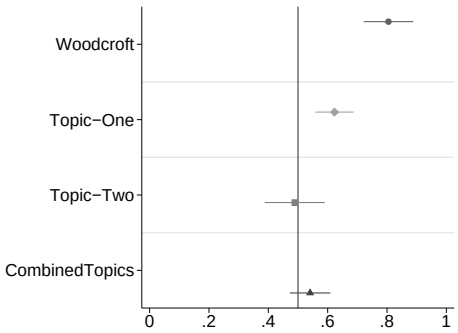
(d) Food patents



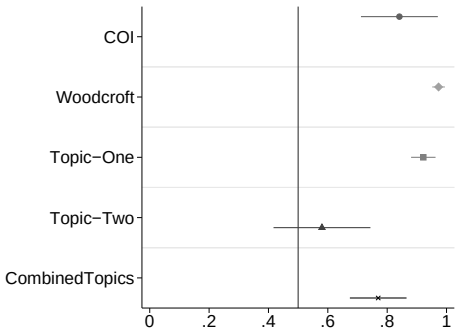
(e) Instrument patents



(f) Medicine patents

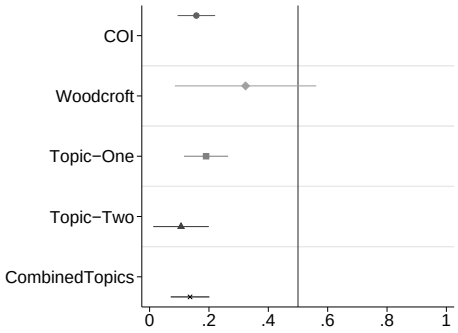


(g) Metal patents

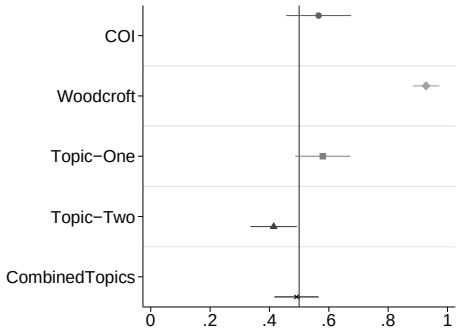


(h) Military patents

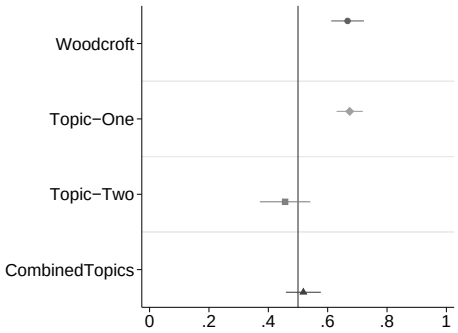
metric, the confidence intervals associated with capital-saving coefficients for all of the common classes are much larger. This may be because of the nature of the variable; 'capital-saving' is a dummy variable, which means the interaction terms with each



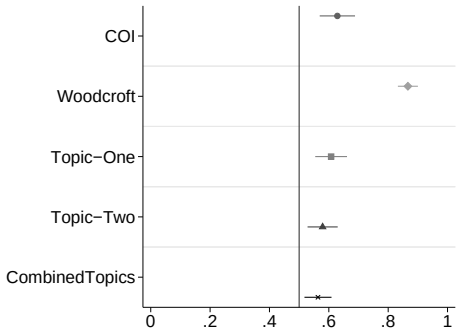
(i) Mining patents



(j) Paper patents



(k) Power patents



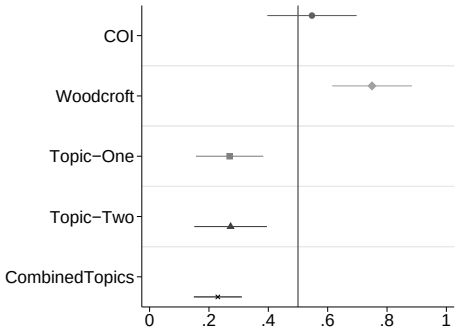
(l) Textile patents

Figure E1: Regression plots for the citations of patented inventions

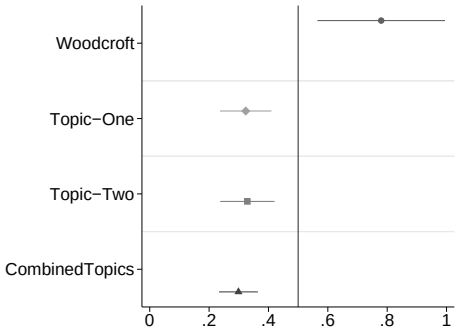
Notes: The plots depict the results from OLS regressions of the Nuvolari-Tartari (NT) schema against the Cradle of Invention (COI), Woodcroft, Topic-One, Topic-Two, and CombinedTopics schemas. The variable of interest is the Woodcroft Reference Index. Each plot represents one of the 12 common patent classes analysed in section 5 of the main text. A value of zero indicates no correlation between schemas. A value of one should indicate complete correlation between schemas, suggesting both taxonomies classify patents in the same way. The confidence intervals for coefficients reflect the fluctuations in terms of significance, and the position of coefficients represents the fluctuations in terms of size.

patent class are also dummies, while patent quality is a continuous numerical variable. Consequently, standard errors may be much larger as we are dealing only with values of zero and one. In addition, the capital-saving metric is the only metric which reports both complete and zero correlation coefficients For Clothing patents, in sub-figure E2c, the Topic-Two schema reports no correlation with the NT schema, which suggests that they have captured completely different patents. While Medicine patents, E2f, report a coefficient of zero for Topic-One, and a coefficient of one for Woodcroft. This indicates significant disparities between schemas, as Woodcroft and NT seem to classify the same patents, while Topic-One does not capture any similarity at all.

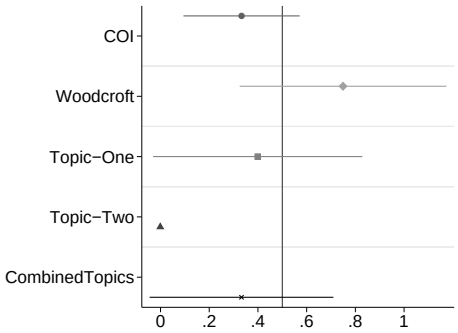
Comparing these results with the regressions in Table 9 may highlight whether the regression plots could predict the likely degree of classification divergence. Similar to



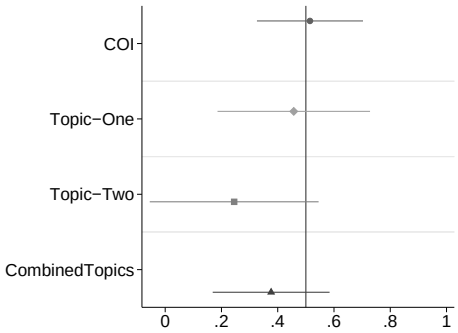
(a) Agricultural patents



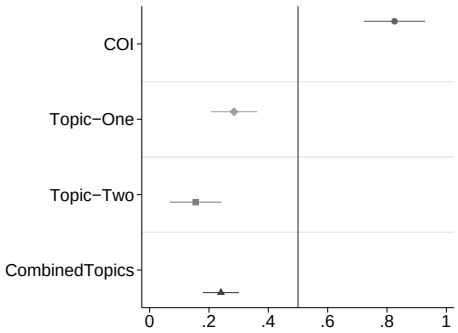
(b) Chemical patents



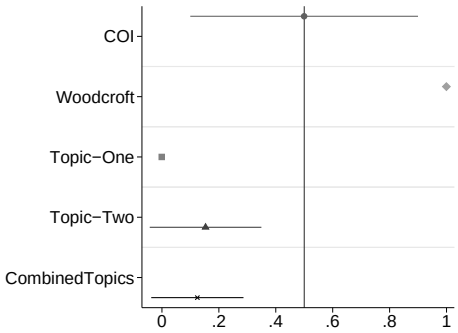
(c) Clothing patents



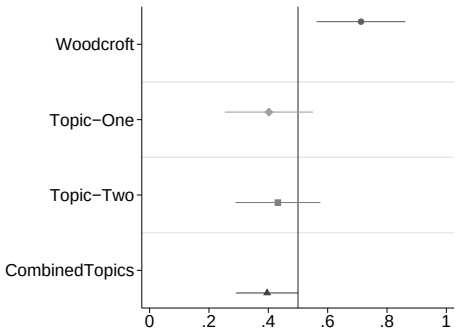
(d) Food patents



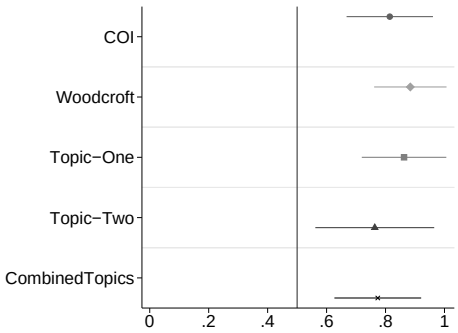
(e) Instrument patents



(f) Medicine patents

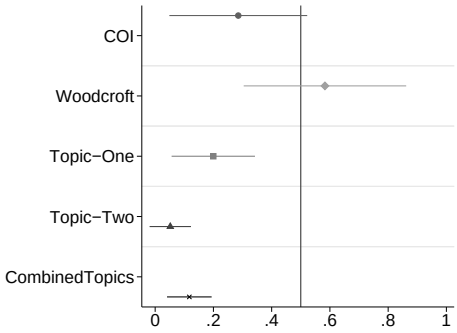


(g) Metal patents

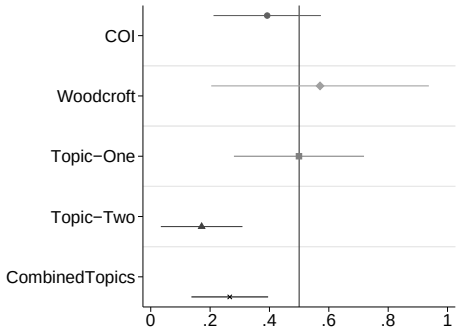


(h) Military patents

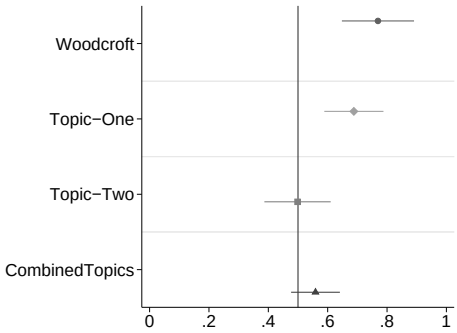
patent quality, the COI and Woodcroft schemas are correlated with NT, so we may expect their results to be similar when compared the machine learning schemas. For example, the Woodcroft schema is strongly correlated with NT for both Agricultural and Chemical



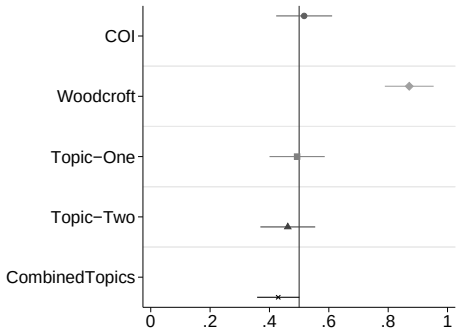
(i) Mining patents



(j) Paper patents



(k) Power patents



(l) Textile patents

Figure E2: Regression plots for capital saving patents

Notes: The plots depict the results from OLS regressions of the Nuvolari-Tartari (NT) schema against the Cradle of Invention (COI), Woodcroft, Topic-One, Topic-Two, and CombinedTopics schemas. The variable of interest is a dummy variable indicating whether a patent is capital-saving. Each plot represents one of the 12 common patent classes analysed in section 5 of the main text. A value of zero indicates no correlation between schemas. A value of one should indicate complete correlation between schemas, suggesting both taxonomies classify patents in the same way. The confidence intervals for coefficients reflect the fluctuations in terms of significance, and the position of coefficients represents the fluctuations in terms of size.

patents, therefore we may expect to see similar results for Chemical patents in Table 9 for both of those schemas. Indeed, this is what we observe, as both Woodcroft and NT report similarly sized coefficients, albeit with some difference in statistical significance. The regression plots for Metal patents, in sub-figure E2g also show the Woodcroft schema is more correlated with the NT schema than any of our schemas. But, in Table 9 the Topic-One schema reports a result much closer to the NT schema's, even though they are less correlated in our regression plots. Overall, this suggests that the regression plot method cannot reliably be used to predict classification divergence outcomes.

Appendix E.3 – Discussion

The regression plots describe the degree of correlation between the NT schema and each of the alternative patent taxonomies. By creating a series of interaction terms between each of our six patent metrics and each of the alternative taxonomies, we could then regress each interacted taxonomy against the interacted NT schema for each metric. Regressing each interacted schema independently allows us to identify how correlated those patent schemas are, which is useful for understanding how similarly they classify patents.

Across all six metrics, the degree of correlation is very similar. Generally, the COI and Woodcroft schemas are found to be more strongly correlated with the NT schema than Topic-One, Topic-Two, or CombinedTopics. This may indicate that our machine learning classifies patents in a considerably different way compared with those schemas which rely more extensively on manual classification. This is not to say our methodology is 'correct', but rather to point out that the differences could be categorised as machine versus human judgement. Because of the strong similarities in terms of coefficient size, we opted to report the results only for the patent quality and capital-saving metrics. The major difference arising from observing the capital-saving metric is the size of the confidence intervals for all coefficients, which were significantly larger than those observed in relation to the patent quality metric.

The variation observed in section 5 in the main paper coupled with the strong similarities for the regression plots suggest that we cannot use taxonomy correlations to predict the likely degree of classification divergence. The capital-saving metric, for example, reports the greatest degree of classification divergence in our main results. But, capital-saving regression plots are strongly similar to patent quality regression plots, apart from the differences to confidence intervals. The larger confidence intervals are unlikely to explain the divergence in terms of coefficient size or direction of association.

Overall, we find that the regression plot method is not a reliable means for predicting classification divergence outcomes. For all six metrics, there are too many dissimilarities for this method to predict any outcomes with confidence. As shown for the patent

quality and capital-saving metrics, there are instances where schemas that are correlated produce similar divergence results, but there are also instances where the opposite is true, Consequently, the regression plot results are useful for a general understanding of how patent taxonomies are correlated, but they provide no predictive power for identifying possible outcomes in relation to classification divergence.

References

A Cradle of Inventions: British Patents from 1617 to 1894 (2009), *Stevenage, UK: Metal Finishing Information Services Ltd.*

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), 'Latent dirichlet allocation,' *Journal of machine Learning research,* ***3***, 993–1022.

Bottomley, S. (2014b), *The British patent system during the Industrial Revolution 1700-1852: From privilege to property.* Cambridge University Press.

Carpmael, W. (1842), *The law of patents for inventions, familiarly explained for the use of inventors and patentees.* London: Simpkin, Marshall, and Co., Stationer's-Hall Court; and Weale, High Holborn.

Chen, Y., H. Zhang, R. Liu, Z. Ye, and J. Lin (2019), 'Experimental explorations on short text topic mining between lda and nmf based schemes,' *Knowledge-Based Systems,* ***163***, 1–13.

Dutton, H. I. (1984), *The patent system and inventive activity during the Industrial Revolution, 1750-1852.* Manchester University Press.

Hutchins, L. N., S. M. Murphy, P. Singh, and J. H. Graber (2008), 'Position-dependent motif characterization using non-negative matrix factorization,' *Bioinformatics,* ***24***(23), 2684–2690.

Jewkes, J., D. Sawers, and R. Stillerman (1969), *The Sources of Invention* (2nd ed.). Macmillan, London.

Jones, B. (2009), 'The burden of knowledge and the "Death of the Renaissance Man": Is innovation getting harder?,' *The Review of Economic Studies,* ***76***1, 283–317.

Khan, B. Z. (2015b), 'The impact of war on resource allocation: "Creative Destruction," patenting, and the American Civil War,' *Journal of Interdisciplinary History,* ***46***(3), 315–353.

Khan, B. Z. (2018), 'Human capital, knowledge and economic development: Evidence from the British Industrial Revolution, 1750–1930,' *Cliometrica,* ***12***(2), 313–341.

Khan, B. Z. and K. L. Sokoff (2001), 'The early development of intellectual property institutions in the United States,' *Journal of Economic Perspectives,* ***15***(3), 233–246.

Khan, B. Z. and K. L. Sokoloff (2004), 'Institutions and democratic invention in 19th-Century America: Evidence from 'great inventors,' 1790-1930,' *National Bureau of Economic Research Working Paper Series No. 10966.*

Klemp, M. and J. Weisdorf (2012), 'The lasting damage to mortality of early-life adversity: Evidence from the English famine of the late 1720s,' *European Review of Economic History,* ***16***(3), 233–246.

MacLeod, C. (2002), *Inventing the Industrial Revolution: The English patent system, 1660-1800.* Cambridge University Press.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011), 'Optimizing semantic coherence in topic models,' In *Proceedings of the conference on empirical methods in natural language processing*, pp. 262–272. Association for Computational Linguistics.

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

Moser, P. (2012), 'Innovation without patents: Evidence from World's Fairs,' *The Journal of Law & Economics,* ***55***(1), 43–74.

Nicholas, T. (2010), 'The role of independent invention in U.S. technological development, 1880-1930,' *The Journal of Economic History,* ***70***(1), 57–82.

Nicholas, T. (2011b), 'Independent invention during the rise of the corporate economy in Britain and Japan,' *Economic History Review,* ***64***(3), 995–1023.

Nuvolari, A. and V. Tartari (2011), 'Bennet Woodcroft and the value of English patents, 1617-1841,' *Explorations in Economic History,* ***48***(1), 97–115.

O'Brien, P. K., T. Griffiths, and P. A. Hunt (1996), *Technological change during the first industrial revolution: the paradigm case of textiles, 1688-1851.* Routledge.

O'Callaghan, D., D. Greene, J. Carthy, and P. Cunningham (2015), 'An analysis of the coherence of descriptors in topic modeling,' *Expert Systems with Applications,* ***42***(13), 5645–5657.

Schmookler, J. (1966), *Invention and economic growth.* Harvard University Press.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler (2012), 'Exploring topic coherence over many models and many topics,' In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961. Association for Computational Linguistics.

Van Leeuwen, M. H. and I. Maas (2011), *HISCLASS: A historical international social class scheme.* Leuven University Press.

Woodcroft, B. (1860), *Subject-matter index of patents of invention, from March 2, 1617 (14 James I.) to October 1, 1852 (16 Victoria).* London: Queen's Printing Office.

Wuchty, S., B. F. Jones, and B. Uzzi (2007), 'The increasing dominance of teams in production of knowledge,' *Science,* ***316***(5827), 1036–1039.