



That's Classified! Inventing a New Patent Taxonomy

Billington, S., & Hanna, A. (2020). That's Classified! Inventing a New Patent Taxonomy. *Industrial and Corporate Change*, 30(3), 678-705. <https://doi.org/10.1093/icc/dtaa049>

[Link to publication record in Ulster University Research Portal](#)

Published in:
Industrial and Corporate Change

Publication Status:
Published (in print/issue): 27/12/2020

DOI:
[10.1093/icc/dtaa049](https://doi.org/10.1093/icc/dtaa049)

Document Version
Author Accepted version

General rights
Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk.

That's Classified!

Inventing a New Patent Taxonomy

Stephen D. Billington[†] Alan J. Hanna[‡]

July 2020

Abstract

Innovation researchers currently make use of various patent classification schemas, which are hard to replicate. Using machine learning techniques, we construct a transparent, replicable and adaptable patent taxonomy, and a new automated methodology for classifying patents. We contrast our new schema with existing ones using a long-run historical patent dataset. We find quantitative analyses of patent characteristics are sensitive to the choice of classification; our interpretation of regression coefficients is schema-dependant. We suggest much of the innovation literature should be carefully interpreted in light of our findings.

Keywords: Innovation, Invention, Machine Learning, Patents, Patent Classification, Taxonomy, Economic History.

JEL Codes: K11, N24, N74, 031, 033.

[†]Corresponding author: Department of Accounting, Finance, and Economics, Ulster University. Email: s.billington@ulster.ac.uk.

[‡]Queen's Management School, Queen's University Belfast. Email: a.hanna@qub.ac.uk.

1 Introduction

Patent statistics are a widely used proxy for measuring and understanding technological change (Griliches, 1990). Patents are particularly popular in the field of economic history where alternative measures of inventive activity are few. Patentable inventions, however, have heterogeneous characteristics which, if not accurately controlled for, have the potential to influence any interpretation of patent statistics. For example, the propensity to patent varies by industry, suggesting the decision to obtain a patent also varies by industry (Cohen et al., 2000).

Classification systems can account for varying patent characteristics. Innovation scholars and social scientists can adopt one of two classification approaches: a taxonomy of their own construction; or patent office classifications, such as the International Patent Classification (IPC), United States Patent Classification (USPC), or Cooperative Patent Classification (CPC). The prevalence of both approaches in the literature raises the following questions: how comparable are existing studies that use different taxonomies and methods, and which, if any, of the existing taxonomies can and should be used in future studies?

The use of both manually constructed taxonomies as well as patent office schemas complicates our ability to interpret the results of patent studies. Manual classification is subjective, and there is no guarantee any two patent studies would classify the patents in the same way. Few taxonomies see repeated usage; academic schemas are often not fully discussed or described, hindering successful replicability as well as adaptability. By contrast, patent office schemas often see repeated usage, but are prone to problems of scope, as they usually consist of a few broad categories encompassing thousands of narrowly defined subclasses. Not only are narrow subclasses too fine-grained for econometric usage (Benner and Waldfogel, 2008), but the broad classes suffer problems associated with the purpose of the schema. This problem was first identified by Schmookler (1966) who observed patents for a toothpaste tube and a manure spreader to fall into the same class, simply because they both had the same function of ‘dispensing solids’ (Schmookler, 1966, p. 20). Recently, innovation scholars have made

similar comments (e.g., McNamee, 2013), stating ‘the best way forward is to devise identification systems that avoid entirely technology classification systems that were devised for the sole purpose of helping examiners locate prior art’ (Thompson and Fox-Kean, 2005, p. 466).

Patents can be classified under “time-invariant” or “time-variant” schemas. Time-invariant schemas consist of broad classes, reflecting industry lines, which do not change over time. By comparison, time-variant schemas encompass fine-grained classes, reflecting time- or country-specific innovation; patent office schemas are time-variant since they are used by patent examiners to identify prior art for new patent applications. This relies on very narrowly defined technology boundaries which must evolve as newer technologies emerge that defy classification (Strumsky et al., 2012). Time-variant schemas can also have broad classes, but these are composed of technology subclasses related by technical function for patent examiner convenience. By contrast, time-invariant taxonomies are useful for relating patenting to economic phenomena over time and space, and in the long-run.

Modern patent systems do not vary much, having largely homogenised over time. The long-run, however, encompasses numerous instances of patent reform. Such events act as natural experiments, which can provide important insights concerning the optimal design of patent institutions. The ability to contrast patents throughout history is important for developing a fuller understanding of how patent systems affect innovation, and how they have evolved. Innovation historians make extensive use of time-invariant schemas. Of course, a “one-size-fits-all” schema may not be useful or appropriate for answering all possible research questions. Patent researchers require a standardised schema that can be adapted to suit specific research questions while ensuring patents continue to be classified consistently.

We present a new, time-invariant patent taxonomy for producing more consistent and comparable results within the innovation literature. Our taxonomy is built on the principle of transparency, so future investigators can understand its design and application. In this way, investigators can either: reuse our schema, adapt it for their

own needs, or even develop new schemas using our methods. We also propose to use machine learning techniques for automating patent classification. Machine learning techniques minimize the subjective element of classification, reducing the probability of patents being classified inconsistently. A consistent approach to patent classification will lead to increased comparability of innovation studies, which helps to strengthen our understanding of innovation and its relationship with patents. In turn, this can only benefit policymakers in designing appropriate measures to encourage innovation.

The main focus of this paper is concerned with developing a new taxonomy, and an automated method for classifying patent data. We use text as data to derive our set of time-invariant patent classes. Because the literature abounds of alternative patent taxonomies, we can observe which classes appear frequently. Frequent classes in historical taxonomies reflect technology groups independent of time or geography, making them indicative of time-invariant classes. We then test the validity of our schema by applying machine learning techniques to multiple patent datasets. Patent data have historically contained rich textual information in their titles. Using these titles, we elicit a set of common word associations, or “topics”. Topics capture related patents into technology groups, and can be used to observe whether we have omitted any potential classes. Finally, we use our machine learning techniques to automate patent classification.

We further investigate whether the choice of schema influences the results of examining patent characteristics. To test this, we examine the population of British patents granted between 1700 and 1850. There are several advantages to using this dataset. First, we can reproduce several taxonomies that have classified the data (e.g Woodcroft, 1860; *A Cradle of Inventions*, 2009; Nuvolari and Tartari, 2011). Second, Britain’s patent system remains relatively unchanged during this time (Dutton, 1984), allowing us to better investigate the nature of invention over the long-run. Third, the data span the Industrial Revolution – amongst the most important phenomena in human history (Clark, 2007). Insights from this period are helpful to address important questions about the origins of the wealth of nations and the role of innovation in industrialisation.

We contribute to the field of innovation studies through the creation of a new, well-

defined, time-invariant patent taxonomy. We thoroughly describe the development of our schema to ensure future users understand how it was constructed and how it can be adapted or adopted to compliment existing work. We also provide a new methodology for automating the classification of any patent dataset. Our method can lead to more consistent classification with fewer errors, and decreases the time needed to classify large patent datasets, facilitating large-scale patent analyses.

We also contribute by documenting the existence of “classification divergence” – statistical significance, direction of association, and coefficient magnitude are all subject to the choice of classification used in a regression analysis of patent characteristics. To our knowledge, we are the first to identify this kind of sensitivity in the innovation literature, and its implications could be severe. Within the innovation history literature, divergence complicates our understanding of important historical events related to inventive activity, such as the Industrial Revolution. Divergence also makes it difficult for policymakers to develop measures to encourage innovation based on existing findings.

We also contribute to the growing trend in applying text analytics to patent data. With the development of more sophisticated text analysis techniques, the rich textual data contained in patents can be better exploited to advance our understanding of patent systems. Textual analysis usually centres around quantifying the text contained in patents to develop ‘similarity’ measures: pairwise correlations of patents based on their text (Younge and Kuhn, 2016; Arts et al., 2018). Such approaches have strong applications to developing new measures of innovativeness (Kelly et al., 2018), identifying knowledge spillover effects (Feng, 2019; Blit and Packalen, 2019), understanding stock-return variability (Khimich and Bekkerman, 2016), identifying general or specific purpose technologies (Packalen and Bhattacharya, 2012), and measures of patent novelty (Balsmeier et al., 2018).

Our paper is closely related to this recent trend, particularly Younge and Kuhn (2016); Arts et al. (2018); Feng (2019). In their articles, the authors develop measures of patent similarity using common keywords contained in patent specifications or abstracts – the

detailed technical documents outlining what is new about a particular patented invention and how to use it. Such methods are useful for studying innovation but have several drawbacks. First, the articles do not produce long-run classification schemas for use in historical analysis. Second, whether similarity captures related patents based on the patent’s technical function or its intended usage is unclear given the authors rely on the technical patent specifications. Third, similarity measures say nothing about the comparability of existing studies. By contrast, we endeavour to classify patent data according to the industry of final use. In doing so we exclusively use patent titles, which contain fewer references to technicality, and are more likely to help us identify a patent’s relevant industry.

The paper proceeds as follows: Section 2 surveys existing patent classification literature. Section 3 outlines our machine learning techniques. Section 4 details our taxonomy’s construction and validation. Section 5 reports the results of contrasting patent taxonomies in our analysis of patent classes on patent characteristics. Section 6 interprets our findings and discusses their implications . Section 7 concludes.

2 Patent Classification Literature

In 1830, John D. Craig, the US Superintendent of Patents, gave evidence to the US House of Representatives regarding the development of the US patent classification schema. In his evidence, Craig raised two points: the ‘imperceptible shades of difference’ of patent classes, and that ‘a doubt frequently arose concerning the class to which some of the patents did properly belong’ (cited in Bailey, 1946, p. 466). Craig was concerned with the overlapping characteristics of patented inventions. Accurately pinpointing a particular class for a particular patent is a difficult task.

It is precisely because assigning classes is difficult that a standardised schema and classification methodology is necessary. The likelihood of inconsistent patent classification increases without a standard approach. Pearce (1957, p. 1) discusses how the inconsistent classification of industrial statistics led to any comparisons of industrial data as ‘difficult

and often misleading’, resulting in the creation of the Standard Industrial Classification (SIC) schema.

This problem persists in the innovation history literature, particularly as patent data are a popular proxy to study inventive behaviour over the long-run. Innovation historians desire a classification approach which classifies patent data according to industry of final use, as this allows them to study patenting and the wider economy. At present, different authors classify (often the same) patent data in different ways. Patent classes should account for common characteristics that could influence how we interpret such data. If scholars do not adopt the same approach to identifying group-specific characteristics, then studies cannot be easily compared.

2.1 Patent Office Schemas

The development of patent office schemas is primarily to aid patent examiners (WIPO, 1992). Examining patents requires multiple well-trained, professional patent examiners to engage in the time-consuming search for prior art – old or existing patents likely to influence or anticipate future ones. Assigned classes are subject to change, as patents are examined at several different stages throughout the patent application (Righi and Simcoe, 2019). Having thousands of well-defined classes and subclasses facilitates a more efficient search for prior art. Such classes make fine distinctions between seemingly similar types of inventions, allowing examiners to find the relevant art more readily.

Patent office schemas are most suitable for time-variant classification because they primarily classify patents by technical functions. Such methods are extremely useful for identifying emerging technologies (Strumsky et al., 2012) and cumulative innovation (Boschma et al., 2014; Rigby, 2015; Youn et al., 2015) and also for observing the increasing modularisation of technology (Arts and Veugelers, 2015). However, these methods are less useful for relating patenting and the real economy. Despite this, there have been growing attempts to utilise time-variant patent office schemas more widely.

Concordances between patent office schemas and industrial classification schemas have become a popular method to relate patenting to the wider economy (e.g. Verspagen et al.,

1994; Kortum and Putnam, 1997; Johnson, 2002; Schmoch et al., 2003; Costantini et al., 2015; Leydesdorff et al., 2017). This approach links industrial classes to IPC subclasses, either manually (Schmoch, 2008) or through probabilistic algorithms matching keywords from industrial schemas to patent titles (Lybbert and Zolas, 2014). Recent efforts by the United States Patent and Trademark Office (USPTO) have matched USPC classes to the popularly used NBER patent classes of Hall et al. (2001) to facilitate long-run analyses of US patenting behaviour (Marco et al., 2015).

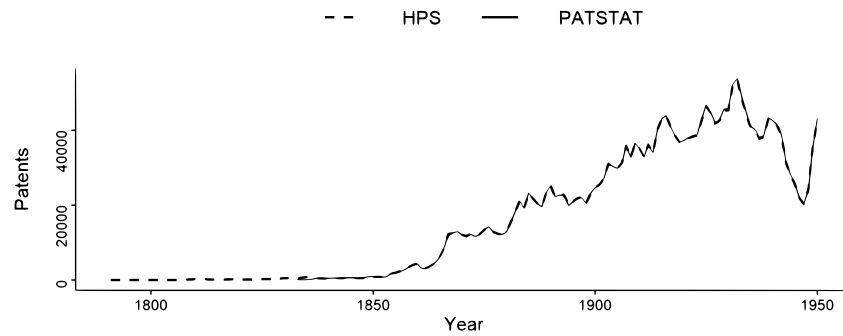
This sub-field, however, relies on existing patent office schemas applied to the data. For the economic historian concerned with historical patent data, the approach is not always feasible. Many historical patents do not have patent office classes assigned, and such classes cannot be easily assigned by academics themselves because of the highly-skilled and specialised nature of patent office classification. The exception here is the case of Marco et al. (2015), while their approach is useful for innovation historians, it is exclusive to US patents.

Figure 1 provides a comparison between the European Patent Office's (EPO) PATSTAT Biblio (Autumn 2016) database and an external count of patent grants pre-1950, for Britain, France, Germany and the United States.¹ PATSTAT claims to contain all patents ever granted in EPO countries, and is a popular data source for innovation studies.² However, pre-1900 patent records are incomplete in PATSTAT; patent office classes are usually unavailable. As the figure shows, PATSTAT is not a suitable data source for historical patent data, and therefore patent office schemas cannot be used.

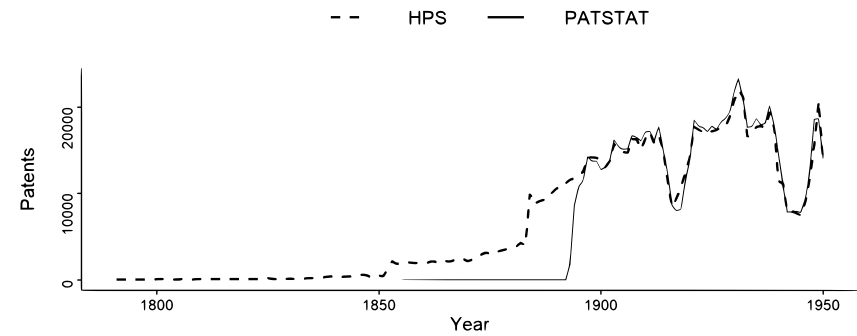
Aside from omitted data, academic studies face further problems using patent office schemas because they classify patents based on two separate principles: "application-oriented" and "function-oriented". The former refers to 'a thing "in general", i.e., characterised by its intrinsic nature or function' (WIPO, 2016, p. 22); the latter refers to 'a thing "specifically adapted for" a particular use of purpose' and 'the incorporation of a thing into a larger system', where a thing refers to any technical

¹Countries chosen due to their popularity in the economic history literature.

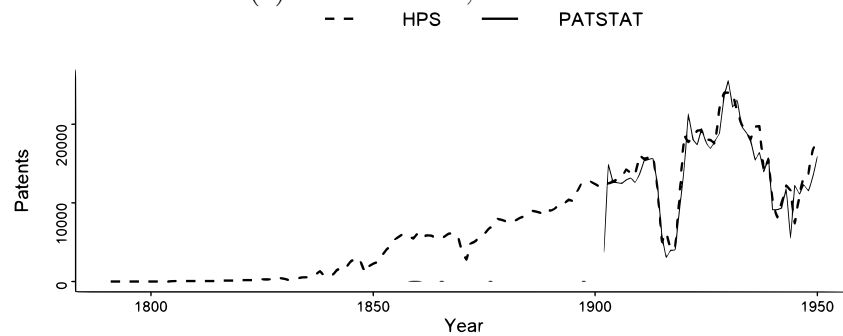
²See the online appendix for further information concerning PATSTAT.



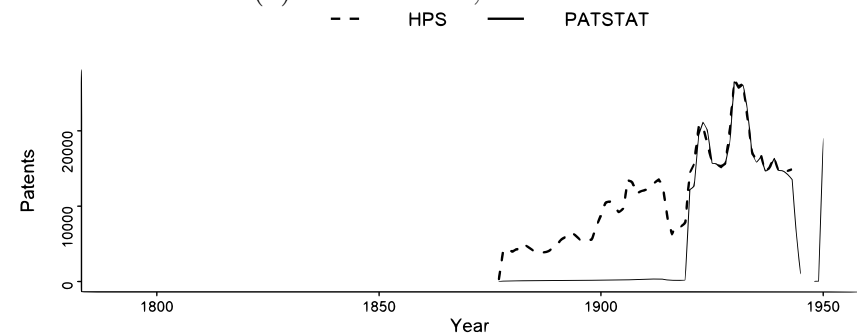
(a) *United States, 1791-1950*



(b) *Great Britain, 1791-1950*



(c) *France, 1791-1950*



(d) *Germany, 1877-1950*

Figure 1: *PATSTAT versus Historical Patent Counts, 1791-1950*

Notes: Comparison of the counts of patents granted in PATSTAT against the Historical Patent Counts data for the US, Great Britain, France, and Germany. Germany's series begins in 1877 since this is when Germany was established.

Source: PATSTAT Biblio Autumn Edition (2016) and Frederico (1964).

subject matter.

Both principles are seemingly jointly applied within the EPO and other patent examination bodies. While the application-oriented approach is more useful for academics, the use of the function-oriented principle suggests patents are being grouped by a function which could apply to multiple industries. Since scholars frequently study the wider effects of patenting, the use of function-oriented classification hinders the usefulness of patent office schemas for this purpose.

Patent offices are becoming more aware of this hindrance, as patent examination bodies are currently adapting their patent classification systems to better capture their final economic use in new products which combine different technologies. One example of this is the EPO's work on the so-called "Internet of Things" (EPO, 2017); another is the work undertaken at the USPTO (Marco et al., 2015). These adaptations, however, continue to rely on the function-oriented approach. In the EPO's case, for example, examiners produced a concordance between CPC classes and a 'cartography' for identifying "Fourth Industrial Revolution" (4IR) technologies. But identifying 4IR patent applications relies on an examination of a patent's CPC codes by patent examiners, who then recommend which field of the cartography the patent belongs to.

2.2 Classification in Economic History

Patent classification in innovation studies allows researchers to relate patent data to changes in the economy. For economic historians studying innovation, this necessitates taxonomies which cover the long-run: time-invariant schemas. Time-invariant schemas group related inventions, independent of time or country or specific technical functions for the purpose of economic enquiry. In doing so, researchers are more capable of identifying and controlling for broad characteristics of different types of inventions, which simplifies their study, preventing these characteristics from complicating insights drawn from patent data. For example, mechanical inventions are more likely to be patented than chemical inventions across time and countries, as machines are more susceptible to reverse engineering (Moser, 2005). The necessity of using time-invariant

schemas are to control for such unchanging and broad characteristics, which requires a standardised set of patent classes.

Schmookler (1966) was amongst the first scholars to highlight the fundamental flaw of using patent office classifications for studying the wider economy. Schmookler's observation that unrelated inventions can be grouped together by a patent office, because of a related functionality, meant social scientists and innovation historians would need to adopt new taxonomies directly related to industry. One of Schmookler's key innovations was to overtly classify patents based on their industry of final use.

In his seminal article, Griliches (1990) builds on Schmookler's efforts, and outlines three methods of classification: "Origin", "Production", and "Destination". Origin groups patents by the industry that produced them. Production groups patents by the industry most likely to produce the invention, or use within the production of goods or services. Destination groups patents according to the industry most likely to make use of, or benefit from, the invention. Destination and Production overlap. The major difference is the use of an invention does not intrinsically imply its use is in production, but an invention used within production constitutes Destination. Destination is the most suitable method for studying patenting within the wider economy, as it is easier to determine a patent's intended industry of final use.

However, developing a single standardised, replicable and adaptable taxonomy has received little attention in the literature. Most studies adopt a set of single-use classes as controls for their econometrics. What is unclear is how authors have constructed their taxonomies, defined their classes, and assigned patents to those classes. Without this information, replicating existing taxonomies is difficult. This leads to two possible outcomes. Either future investigators apply existing taxonomies incorrectly – leading to further inconsistencies – or they construct additional taxonomies, further compounding the problem of incomparability.

Table 1 displays a set of 35 patent and invention taxonomies, which have been used within the economic history literature.³ This set is representative of the economic history

³The set includes papers which make use of Exhibition invention classes, which are then matched up with patent classes. Exhibition classes operate similarly to patent ones, and are therefore included.

literature; the papers were compiled from a systematic search of “economic history” and “patent” keywords in the Econlit and EBSCO databases. We focused on papers observing patent data on the wider economy, and not specific industries, in an attempt to avoid very narrow research questions requiring very specific classification. Furthermore, we only included studies which published their taxonomy, or made reference to another published taxonomy. The table lists the number of classes in each paper’s taxonomy, the countries observed, the time-period examined, which of the three methods of classification were used (if any), and a brief description concerning the construction of the taxonomy and whether definitions were provided for patent classes.

The table highlights several things. First, the existing patent evidence derives from a number of countries (14 in total) and a number of different time periods. This highlights the need for a time-invariant taxonomy to classify similar patents in the same way, independent of location or time. Second, the Destination method of classification is the most popular. Most studies explicitly state their classification is based on the industry of final use, as this is most relevant for relating patents and the real economy. We should therefore endeavour to construct a taxonomy based on this method. Third, there is great disparity in the number of classes per taxonomy.

One important question Table 1 raises is how do authors choose the number of classes in their taxonomy? Another question is how do authors classify their patent data once they have chosen the structure of their taxonomy? To understand this, each author would need to detail exactly how they derived and defined their classes, and how they assigned patents to them. For most studies (see ‘Description’ column), a discussion concerning taxonomy construction is not provided, and in almost all cases patent class definitions are not reported. Some authors rely on restructuring existing taxonomies (e.g Moser, 2005; Brunt et al., 2012; Moser, 2012); or use concordances between patent schemas and industrial schemas (e.g. Nanda and Nicholas, 2014); or IPC headings (e.g. Sáiz, 2014). In most cases, authors construct new taxonomies without description or explanation. Future researchers therefore gain limited insights into how to design a taxonomy, how to classify patents consistently, or how to choose the number of classes for their taxonomy.

Table 1: Classification Literature

Authors	Classes	Country	Time-period	Method	Brief Description
Woodcroft (1860)	246	England	1617-1852	Destination	The author compiled a complete collection of all patents granted in England prior to 1852. The associated taxonomy was most likely to reduce search costs for previously granted patents, akin to schemas constructed by and for patent examiners. Each broad class has an associated list of keywords.
Sokoloff (1988)	4	US	1790-1846	Destination	The author likely uses the NBER patent data schema, although it is not explicitly stated. If not, how the taxonomy was designed is unclear. Class definitions are not provided.
Sullivan (1990)	7	England	1711-1850	Destination	The author constructs their taxonomy to study six important industries. It is unclear how patents are classified, although this is based on Woodcroft's Subject-matter index. Class definitions are not provided.
Basberg (1997)	21	Norway	1839-1860	Destination	The author constructs their own schema based on a combination of broad and narrow classes. It is unclear how the taxonomy was constructed nor how patents are classified. Class definitions are not provided.
Magee (1999)	32	Australia	1858-1902	Destination	The author constructs their taxonomy based on the standard classification of manufacturing industries used in Australia in 1902. From this the author constructs 32 classes. Keywords are provided for certain classes. It is unclear how the author assigns patents to classes.
Cantwell (2000)	15	Britain, France, Germany, Sweden, Switzerland, US	1920-1995	Unclear	The author constructs their taxonomy based on the USPTO classification schema. It is unclear exactly how the author has derived their classes nor how patents are classified. Class definitions are not provided.
Dauids (2000)	16	Dutch Republic	1580-1720	Unclear	The author constructs their own schema. It is unclear how the taxonomy is constructed nor how patents are classified. Class definitions are not provided.
Khan (2000)	23	US	1870-1895	Destination	The author constructs their own schema. It is unclear how the taxonomy is constructed nor how patents are classified. Class definitions are not provided.
Moser (2005)	7	England and US	1851, 1876	Unclear	The author uses exhibition data and constructs their taxonomy based on the 1851 Crystal Palace Exhibition classification scheme. It is unclear why the author chose 7 classes or how these class have been constructed. There is no explicit discussion concerning how US inventions have been classified according to the schema.
Basberg (2006)	18	Norway	1860-1914	Unclear	The author constructs their taxonomy from the official patent office classification scheme. It is unclear how their taxonomy has been designed nor how patents are classified. Class definitions are not provided.
Streb et al. (2006)	89	Germany	1877-1918	Production	The author's patent data has been classified according to the historical German Patent Office classification schema comprising 89 technological classes. This approach is more likely to capture patent functionality rather than application
Baten et al. (2007)	19	Germany	1895-1913	Origin	The author relies on classifying patents according to the Standard Industrial Classification (SIC) two-digit code of the firm who held the patent. This method captures the 'Origin' of patents. The research question focuses on industry spillover effects for innovation, so this is likely a reasonable approach.
Nicholas (2008)	4	US	1910-1939	Unclear	The author uses the NBER patent data, and likely constructs their 4 patent classes from this. However, it is not explicitly described how the classes have been constructed; one class is 'Other' which likely captures a number of different technologies. It is unclear why the author has chosen 4 classes instead of the 6 NBER classes, or even the NBER subclasses.
A Cradle of Invention (2009)	15	England	1617-1852	Unclear	The compilers of the data constructed a simple taxonomy to make the database more useful. They assert that their schema is based on their interpretation and so may be flawed. Class definitions are provided.

Continued on next page

Table 1 – continued from previous page

Authors	Classes	Country	Time-period	Method	Brief Description
Greasley and Oxley (2010)	8	New Zealand	1861-1939	Destination	The author's construct their taxonomy to match their commodity output groups. Their taxonomy is based on Magee (1999), but with fewer classes. It is unclear how patents are classified. Class definitions are provided.
Meisenzahl and Mokyř (2011)	12	England	1660-1830	Unclear	The author's construct their own schema. It is unclear how the taxonomy has been designed nor how patents are classified. Class definitions are not provided.
Nicholas (2011c)	30	Britain, Germany, Japan, US	1900-1940	Production	The author classifies their patents using 30 main categories of the International Patent Classification (IPC) taxonomy. As described in the text, the IPC method classifies patents based on functionality, and so may be classifying un-related technologies as similar.
Nicholas (2011a)	16	US	1921-1938	Origin	The author classifies their patents according to the two-digit SIC code for the firm who held the patent. This method captures a patent's origin.
Nuvolari and Tartari (2011)	21	England	1617-1841	Destination	The author's construct their taxonomy based on a working paper version of Moser (2012). It is unclear how they derived their additional classes nor how patents are classified. Class definitions are not provided.
Moser (2012)	10	England and US	1851, 1876, 1893, 1915	Unclear	The author uses exhibition data and constructs their taxonomy based on the 1851 Crystal Palace Exhibition classification scheme. It is unclear how the author derives their classes. Class definitions are not provided.
Brunt et al. (2012)	12	England	1839-1939	Destination	The author's examine prize data from the Royal Agricultural Society of England (RASE) compared with patent data from the British patent office. The author's construct a taxonomy of 12 classes with 130 subclasses based on the Subject-matter Index from Woodcroft (1860), which relies on generating a list of keywords for each patent class. This approach likely relies on a patent's 'destination' for classification.
Burhop and Wolf (2013)	10	Germany	1884-1913	Unclear	The author's report 10 active technology fields, although it is unclear whether there are more. It is also unclear how patents have been classified and according to what kind of taxonomy. Class definitions are not provided.
Khan (2013b)	4	US	1790-1930	Destination	The author reports 4 patent classes in their regression analysis. It is unclear where the classes are derived from, or if any other patent classes are controlled for in their regressions. Class definitions are not provided
Khan (2013a)	12	US	1837-1874	Destination	The author uses exhibition data and patent data. Their taxonomy classifies inventions which were exhibited at industrial fairs. Invention's were classified according to their Destination. It is unclear how the taxonomy was constructed and how inventions are classified. Class definitions are not provided.
Khan (2014)	12	US	1835-1870	Destination	The author uses both exhibition and patent data. They construct their own schema to classify both datasets. It is unclear how the taxonomy has been constructed nor how patents are classified. Class definitions are not provided.
Nanda and Nicholas (2014)	15	US	1921-1938	Origin	The author's classify their patents according to the two-digit SIC code for the firm who held the patent. This method captures a patent's origin.
Khan (2015a)	26/7	Britain, France, US	1754-1852	Unclear	The author surveys prize-giving institutions in Britain, France, and the US. The French exhibition classes for inventions are those classes produced for the Paris exhibition. For Britain, the invention data are classified according to the Royal Society of Arts (RSA) own designated technology categories. No US classification is described. It is unclear whether the inventions in France and Britain are comparable. Class definitions are not provided.
Sáiz (2014)	20	Spain	1820-1930	Destination	The author constructs their schema based on the IPC descriptions and the descriptions in the patent. It is unclear how they constructed their schema nor how patents are classified. Class definitions are not provided.

Continued on next page

Table 1 – continued from previous page

Authors	Classes	Country	Time-period	Method	Brief Description
Nuvolari and Vasta (2015)	14	Italy	1861-1913	Unclear	The author's construct their schema based on the International Standard Industrial Classification (ISIC). It is unclear how the classes have been constructed nor how the patents are classified. Class definitions are not provided.
Khan (2016)	10	France	1791-1855	Destination	The author's patent data are categorised into industry of final use. It is unclear how the taxonomy has been produced nor how patents are classified. Class definitions are not provided.
Lehmann-Hasemeyer and Streb (2016)	5	Germany	1892-1913	Origin	The author's classify their patents according to the industry of the firm who held them. It is not clear how the taxonomy is formed or how firms are classified. Class descriptions are not provided.
Akcigit et al. (2017)	63	US	1880-1940	Production	The author's use the United States Patent and Trademark Office (USPTO) classification schema and the NBER aggregate classes. Using the USPTO codes patents are matched to the two-digit and three-digit SIC codes. This methodology more likely captures functionality rather than application.
Comino et al. (2017)	17	Venetian Republic	1474-1550	Production	The author's use the technology classification assigned to those patents by another researcher. The author's manually classified their patents, although it is unclear how patents have been assigned. Class definitions are not provided.
Khan (2017)	6	England	1750-1850	Unclear	The author classifies their exhibition data based on the categories used by the RSA to administer prizes to inventions. It is unclear how these classifies are defined or assigned. Class definitions are not provided.
Donges and Selgert (2019)	29	Germany	1843-1877	Production	The author's construct their schema based on the classification scheme of the Imperial Patent Office in 1877. They construct 30 technology groups to match patent data to industrial data, creating a table describing how their classes are defined. The author's assign their schema based on the patent's technical description.

Notes: The table shows a sample of 35 widely disseminated patent taxonomies. 'Classes' shows the number of reported patent classes in each study. 'Country' refers to the country or countries observed in the study. 'Time-period' refers to the period being studied. 'Method' indicates how the taxonomy has been constructed, this column takes one of four values: 'Destination', 'Origin', 'Production' or 'Unclear'. For a method to be recorded it must be either explicitly stated or based on classifying patents according to the firm's own industry (the origin approach). The patent taxonomies highlighted in bold are those used in this study to examine whether the choice of patent classification affects the results from a regression of patent characteristics.

Sources: see Column 1

This is not to say that schemas have been constructed arbitrarily. On the contrary, we acknowledge that existing taxonomies are the result of a great deal of thought and effort, and that authors will have selected or developed a schema to answer a specific research question. The major difficulty, however, is understanding how authors have classified their patent schemas. Without an understanding of how schemas are constructed or how classes are defined, the literature is at risk of classifying similar patents in a number of different and possibly inconsistent ways. Whether this inconsistency matters for interpreting patent data is explored later in this paper.

3 Text Analysis

Text analysis techniques create structured data from natural language documents. We apply these techniques to existing patent classification taxonomies to aid the development of our time-invariant schema. The text describing existing systems is parsed, stemmed and stripped of stop words for analysis.⁴ We subsequently use further techniques to validate the schema, and to assign patents to appropriate classes.

Historically, patent applications contained short, descriptive titles. Under the European Patent Convention, application titles are accuracy checked by patent examiners (EPO, 2017) who can amend the title as they see fit. Under Britain’s historical patent system, patents could be annulled if the invention was not properly described leading to a standardised approach for writing titles (Dutton, 1984; MacLeod, 2002; Bottomley, 2014b). Therefore, historical patent titles provide a rich source of textual data that can be used to classify patents according to their industry of final use.

This study relies exclusively on patent titles, rather than the more commonly used abstracts of patent specifications (see Magerman et al. (2015); Ruckman and McCarthy (2016); Younge and Kuhn (2016); Arts et al. (2018), for examples). We do so for two reasons. First, many historical patents do not possess digitally available abstract or technical specification data – in many cases they do not exist. Using abstracts and specifications would therefore discriminate heavily against historical patent data. Second,

⁴See online appendix for a description of these terms.

we wish to avoid a classification system based on the patent’s technical subject matter. Abstracts and specifications describe how their invention is to be worked, and what is new compared to the prior art, so they can be easily replicated by those “trained in the art”. This requires considerable technical detail. Historically, patent titles concisely state what the patent application is, making them more reflective of potential industries of final usage.

When working with patent data, topic modelling is an increasingly popular methodology to approach classification (Wu et al., 2010; Kaplan and Vakili, 2012; Venugopalan and Rai, 2015; Ruckman and McCarthy, 2016; Wu et al., 2016; Chen et al., 2017; Suominen et al., 2017). This exploits the tendency of patent titles from particular industries to consist of distinctive keywords. Our preferred method of topic derivation is Non-Negative Matrix Factorization (NMF) using the implementation from scikit-learn (Pedregosa et al., 2011). For a discussion of why NMF was chosen rather than the more commonly used Latent Dirichlet Allocation (LDA), please refer to the online appendix.

4 The Taxonomy

In this section, we begin the design, validation, and description of our new, time-invariant patent taxonomy. The taxonomy should be used to classify patent data according to their destination based solely on the text contained in patent titles. It should contain a set of broad, time-invariant classes; those independent of both time and the country granting the patent rights.

To construct and validate our taxonomy, we take a two-step text analysis approach. First, we identify a set of commonly occurring patent classes. We do this by constructing a set of unique words derived from the titles of patent classes contained in the taxonomies reported in Table 1. The sample is representative of the literature, featuring taxonomies from different regions and time periods. This allows us to identify which words frequently appear in the economic history literature by manually grouping together synonyms and

related terms.⁵ Frequently appearing words in historical patent taxonomies are more likely to reflect broad technology groups rather than niche ones, making them suitable candidates for time-invariant classes.

Second, we apply the NMF method to different datasets from PATSTAT covering different countries and time periods. This robustness check helps to ensure that we have not omitted any classes, nor that we have included any niche classes that do not represent the literature, increasing our confidence that we have identified time-invariant classes. In addition, we use NMF to help construct definitions for our taxonomy. Each topic's words are used to assign it to a patent class and are therefore useful to construct the class description.

The taxonomy itself, however, does not immediately dictate the classification approach; patent class definitions will determine whether our schema classifies patents by technical function or industry of final use. The taxonomy will instead derive a set of broad, adaptable classes which should capture the broad characteristics of patented technologies.

4.1 Constructing the Taxonomy

To derive our base set of classes, or “word-groups”, we first undertake a counting exercise on the existing historical taxonomies contained in Table 1.⁶ This exercise compiles all patent classes from each taxonomy into a single set of unique words. Stop words are removed, and stemming applied to avoid duplicate terms. Each unique word is first tallied, then manually grouped with other related words. This is achieved by treating each unique word as a patent class of its own, and then identifying which broader group of inventions each class should belong to. For example, the word ‘agriculture’ is grouped with the terms ‘seeds’, ‘land’, and ‘cultivate’, which are related to the agricultural industry. Within each word-group, the title derives from the highest tallied term. This acts as a reference point,

⁵Human judgement has the advantage here because we need to be able to identify related words and synonyms. A text analysis technique may not be able to do so, because splitting patent class into unique words strips them of their context or definitions.

⁶We use the term word-groups to distinguish from a finalised set of patent classes, as word-groups reflect possible classes which still require verification.

as it provides a noisy estimate of the most commonly occurring classes throughout the literature.

We have deliberately included both patent office and non-patent office schemas in our methodology. Patent office schemas classify patents based on function-oriented characteristics, while non-patent office schemas use application-oriented characteristics; technology-based versus use-based classification. The class definitions, rather than the schema itself, dictate whether patents are classified according to their technical functions or industry of final use. Patent office classes and subclasses consist of multiple words which identify specific technical functions, while non-patent office schemas rely on a single broad class title to convey the likely industry of final use. Our approach decomposes every patent title into unique words, stripping them of their definition, thereby removing any technology or use-based focus. The decomposition of patent titles also removes the influence of highly technical terms as these cannot be readily grouped with other word-groups, reducing the likelihood of our taxonomy being biased by function-oriented characteristics.

Table 2 reports the result of the counting exercise. Our initial count produced 1,105 unique words, the majority of which appeared only once or twice. The first half of the table shows the unique words with the highest initial tallies. This tally on its own, however, is not informative because it does not yet account for synonyms or related text.

The second column of Table 2 shows the results of our manual grouping. As a provisional check, both authors conducted this exercise independently, twice. Each exercise resulted in a similar set of 24 word-groups. ‘NA’ reflects unique words too vague or too technical to assign to any word-groups.⁷ ‘Aggregate Count’ represents the total tally of all related words within each word-group. This count indicates how frequently each word-group appears. Word-groups with lower tallies are less likely to reflect broad patent classes. ‘Heat’, for example, has the lowest count, as terms related to ‘Heat’ did not often appear as part of any class within our sample; ‘Heat’ may not be a suitable patent class.

⁷To ensure ‘NA’ did not reflect its own word-group, we ran another exercise on the words not assigned to other groups. Neither author could readily identify any further sets of word-groups.

Table 2: Counting Exercise Results

Raw Count		Author's Count		Patent Examiner	
Word-group	Frequency	Word-group	Frequency	Word-group	Frequency
product	47	NA	237	NA	398
metal	38	Hardware	226	Engineering Tools	304
machin	31	Chemicals	218	Manufacturing	288
engin	29	Instruments	211	Transport	151
textil	29	Agriculture	182	Light Industry	150
instrument	28	Textiles	176	Household	143
agricultur	26	Transportation	145	Agriculture	113
chemic	25	Construction	124	Construction	109
equip	25	Goods/Services	117	Organisation	98
mine	24	Machinery	111	Textiles	95
transport	23	Paper	99	Foodstuffs	77
food	23	Power	95	Energy Production	72
electr	22	Manufacturing	79	Construction (resources)	60
machineri	21	Metal	73	Wearables	59
paper	20	Food	70	Chemistry	58
manufactur	19	Health	64	Furniture	52
construct	19	Mining	50	Engineering Components	52
print	18	Utility	47	Metallurgy	51
gas	17	Apparel	43	Mining	49
servic	17	Military	43	Engineering Electrical	48
ship	16	Electricity	40	Medical	47
furnitur	16	Gas	35	Weapons	37
glass	15	Light	22	Engineering Civil	19
		Water	19	Telecoms	18
		Communications	17	Machines	6
		Heat	16	Electrical	5

Notes: The 'Raw Count' columns represent our results from the initial frequency counts for words scoring 15 or greater. 'Author's Count' displays the word-groups from our manual assignment. 'Patent Examiner' displays the word-groups from our experienced patent examiner's second manual assignment attempt. 'NA' denotes words too vague to be grouped into word-groups.

Source: See Table 1.

4.2 Taxonomy Validation

We next conduct a series of robustness checks to verify the validity of our word-groups as representative patent classes. These checks include: inviting a senior international patent examiner to undertake our counting exercise just described; ascertaining the frequency of each word-group in our sample of patent taxonomies; and checking whether different patent datasets can be readily classified using our set of word-groups.

As our first robustness exercise, we invited a patent examiner, with over 30 years of experience of classifying patents and revising official patent classification schemas, to conduct our counting exercise. There are several advantages to soliciting an independent review from a patent examiner. First, they are experts in patent classification; it is their job. Second, they are constantly developing and updating classification methodologies, such as their recent work on classifying technologies relating to the Internet of Things. Third, they are trained to classify patents according to both function-oriented and application-oriented characteristics. Fourth, they understand the terminology contained in patent titles. The examiner was made aware of the purpose of our schema, and instructions were provided for them to complete the exercise. The examiner was not shown the results of our initial counting exercise, and was given only the necessary information so as not to bias their attempt. The examiner undertook our exercise twice, on separate dates receiving feedback only on the number and breadth of the categories.

In their second attempt, the examiner derived 24 word-groups, of which 17 word-groups are similar to ours (see final column of Table 2). By comparing the descriptions provided by the patent examiner with the terms included in each of our word-groups, we could subsume the examiner's remaining word-groups into at least one of ours: 'Construction' and 'Construction (resources)' relates to Construction; 'Electrical' relates to Electricity; 'Energy Production' relates to Engines; 'Light Industry' relates to multiple word-groups; 'Metallurgy' relates to Metal; 'Engineering Civil' relates to Construction; 'Engineering Components' and 'Engineering Tools' are both related to Hardware and Instruments; 'Engineering Electrical' is related to Electricity; 'Furniture' and 'Household' relate to

Table 3: Results from Matching Word-groups to the Literature

Word-groups	Total	Percentage
Machinery	29	82.86
Textiles	29	82.86
Chemicals	27	77.14
Metal	26	74.29
Agriculture	26	74.29
Instruments	23	65.71
Food	23	65.71
Paper	21	60.00
Construction	19	54.29
Mining	17	48.57
Electricity	16	45.71
Transportation	15	42.86
Health	15	42.86
Goods	12	34.29
Apparel	11	31.43
Military	11	31.43
Manufacturing	11	31.43
Engines	10	28.57
Utility	10	28.57
Hardware	7	20.00
Communications	7	20.00
Gas	4	11.43
Water	4	11.43
Heat	3	8.57
Light	3	8.57

Notes: The table shows how often each word-group appears in our sample of 35 taxonomies from Table 1. ‘Percentage’ indicate the percentage of taxonomies each word-group appeared in. Word-groups with higher scores are considered more robust and representative of the literature.

Source: Authors’ calculations using data from Table 1.

Goods/Services; ‘Organisation’ relates to Goods/Services and Paper; and ‘Wearables’ relates to Apparel.

We next check how often our word-groups appeared as a distinct class in the sample literature, highlighting whether we derived a schema representative of the literature and whether our word-group titles are sufficiently broad. We check each word-group against each taxonomy, and then tabulate how often they appear. The results are reported in Table 3. For example, 29 out of the 35 sample taxonomies list Machinery as a distinct class, while Heat appears only three times.

Table 4: *Data Used for Topic Analysis*

Country	Years	Source
England ('EN')	1617-1852	<i>A Cradle of Inventions (2009)</i>
France ('FR')	1855-1938	PATSTAT Biblio
Germany ('DE')	1877-1933	PATSTAT Biblio
Great Britain ('GB')	1899-1913	PATSTAT Biblio
United States ('US')	1790-1900	PATSTAT Biblio

Notes: For France and Germany we took a random sample of patents. For Great Britain and the US, we included all patents available until the specified end date. This means many patents were missing for older periods. However, since we care about deriving word-groups, this omission is not serious as we are still capable of observing patent titles.

Source: See Source column.

As our last verification check, we examine the ability of our word-groups to classify different patent datasets, using the machine learning methodology described in Section 3 and the online appendix. Specifically, we derive topics from a number of PATSTAT datasets, and then match these topics to our word-groups. Despite PATSTAT's incompleteness, it is a useful source for examining late-nineteenth and twentieth-century patents from leading industrialised nations. Table 4 shows the data used to verify our word-groups.

We base the strength of our proposed schema on whether it can suitably classify each topic. We are particularly concerned with the spanning nature of the proposed classes: we wish to assign at least one word-group per topic, and are less concerned with instances where ambiguity arises. Large patent datasets will inevitably contain pioneering and niche inventions which are more difficult to classify. Such outlier patents are unlikely to undermine an entire classification schema. Nevertheless, if significant numbers of patents appear as distinct, unclassifiable topics, then our schema is likely to be undermined.

To check the robustness of our word-groups, we apply NMF to the patent datasets described in Table 4.⁸ For France and Germany, we draw a random sample from each decade. By taking samples, we can ensure each dataset is of a similar manageable size, and we can generate the same number of topics.

To justify including additional classes, we should observe topics which cannot be mapped to existing word-groups. Should such a distinct class exist, then we should

⁸See online appendices for more information.

Table 5: Topic Analysis Results using PATSTAT Datasets

Word-group	EN	FR	DE	GB	US
Agriculture	7	3	1	3	5
Apparel	2	2	2	3	2
Chemicals	15	17	10	5	2
Communications	0	1	0	2	4
Construction	4	2	3	3	2
Electricity	1	6	3	6	5
Engines	8	4	6	6	6
Food	1	3	4	1	0
Gas	1	1	2	2	1
Goods	3	6	5	7	5
Hardware	17	17	20	25	28
Health	1	2	0	1	3
Heat	1	0	0	2	2
Instruments	9	6	12	11	11
Light	3	1	0	2	1
Machinery	4	3	5	8	13
Manufacturing	1	1	0	1	3
Metal	4	4	4	2	5
Military	2	3	0	2	1
Mining	2	1	2	1	0
Paper	5	3	1	6	3
Textiles	18	2	0	4	5
Transportation	7	9	2	8	8
Utility	2	5	3	8	5
Water	2	1	0	1	1

Note: The table shows how many topics could be classified according to our set of 24 word-groups. For France and Germany, their total will not add up to 120 due to spurious word associations. This is a result of the difficulties of stemming French and German patent titles.

observe significant numbers of patents and a distinct language describing them. By extracting more topics from each dataset than word-groups within the proposed schema, we attempt to identify any omitted classes.

Table 5 reports our results. Each column represents one of our datasets from Table 4, and their values indicate how many topics could be assigned to a particular word-group: for example, seven topics out of 120 were assigned to ‘Agriculture’ for the England data. For France and Germany there were more unclassifiable topics due to spurious word associations. For most word-groups, we were able to classify multiple topics across each dataset. Few word-groups were poorly represented, such as ‘Heat’ and ‘Light’, suggesting

these are unsuitable patent classes.

Our topic analysis confirms the proposed schema is sufficient to capture patents from a number of diverse datasets.⁹ We could readily assign each topic to at least one word-groups, supporting the word-groups as time-invariant classes. Similarly, the French and German datasets were able to be classified with our proposed word-groups, further supporting the use of our methods for any patent dataset, irrespective of the language.

4.3 Finalising the Taxonomy

To determine whether our proposed schema constitutes a time-invariant taxonomy, we compile the results from each of our exercises into Table 6. The table is sorted by the column ‘Frequency’ which is a simple score indicating the likelihood of a word-group reflecting a patent class. To derive scores for the Frequency column, “cut-off points” are constructed for each exercise. Cut-off points are subjectively determined, and are intended only to help guide the construction of our patent schema; these are described in the table notes. When reviewing the results of our exercises, we place less weight on the Frequency score if word-groups received scores only from the final exercise.

‘Outcome’ records our final decisions regarding the word-groups. Based on our review, we merge the following classes: ‘Communications’ into ‘Electricity’; ‘Gas’ into ‘Chemicals’ or ‘Utility’; ‘Heat’, ‘Light’ and ‘Water’ into ‘Utility’. For the word-groups ‘Mining’ and ‘Manufacturing’, which had high Frequency scores, we kept these due to their scores in the first three exercises; both word-groups received the entirety of their score from the ‘Topic Datasets’ exercise. Additionally, we amend the titles of certain word-groups to broaden their scope as patent classes: ‘Goods/Services’ is renamed ‘Commodities’ and ‘Engines’ is renamed ‘Power’. Our final schema then comprises 20 time-invariant classes.

Finally, we use the topic analysis techniques to help construct class descriptions. Descriptions should ensure patents are classified by industry of final use, rather than

⁹To prepare the patents for analysis, patent titles were stripped of non-printing characters and stop words. Suitable substitutions are applied to reduce all text to a standard character set. Once the topics are generated, they are then translated into English.

Table 6: *Patent Class Methodology Scores*

Word-group	Tally	Patent Examiner	Literature Validation	Topic Datasets					Frequency	Outcome
				EN	US	GB	FR	DE		
Chemicals	218	1	27	15	2	5	17	10	0	–
Metal	73	1	26	4	5	2	4	4	0	–
Construction	124	1	19	4	2	3	2	3	0	–
Transportation	145	1	15	7	8	8	9	2	0	–
Goods/Services	117	1	12	3	5	7	6	5	0	–
Machinery	111	0	29	4	13	8	3	5	1	–
Textiles	176	1	29	18	5	4	2	0	1	–
Agriculture	182	1	26	7	5	3	3	1	1	–
Instruments	211	0	23	9	11	11	6	12	1	–
Engines	95	0	10	8	6	6	4	6	1	–
Hardware	226	1	7	17	28	25	17	20	1	–
Paper	99	0	21	5	3	6	3	1	2	–
Electricity	40	1	16	1	5	6	6	3	2	–
Apparel	43	0	11	2	2	3	2	2	2	–
Utility	47	0	10	2	5	8	5	3	2	–
Food	70	1	23	1	0	1	3	4	3	–
Health	64	1	15	1	3	1	2	0	3	–
Military	43	1	11	2	1	2	3	0	3	–
Mining	50	1	17	2	0	1	1	2	4	–
Manufacturing	79	1	11	1	3	1	1	0	4	–
Communications	17	1	7	0	4	2	1	0	5	Reassigned to Electricity
Gas	35	0	4	1	1	2	1	2	6	Reassigned to Chemicals and Utility
Heat	16	0	3	1	2	2	0	0	6	Reassigned to Utility
Light	22	0	3	3	1	2	1	0	6	Reassigned to Utility
Water	19	0	4	2	1	1	1	0	7	Reassigned to Utility

Notes: ‘Tally’ refers to our initial derivation of unique word-groups. ‘Patent Examiner’ refers to our patent examiner’s derivation of word-groups, we score a value of 1 if the examiner’s word-group matched ours. ‘Literature Validation’ refers to counting the number of taxonomies in which each of our 24 word-groups appear. ‘Topic Datasets’ shows how often each of our word-groups classified topics in each of our five patent datasets. ‘Frequency’ is a simple count indicating how likely a word-group reflects a patent class: the higher the score, the less likely the group should be kept. Word-groups can only have a maximum Frequency of eight, as there are seven distinct exercises. For a word-group to receive a score from any given exercise, we defined “cut-off” points. The cut-off points are defined as follows: ‘Tally’ below 50; ‘Patent Examiner’ score of zero; ‘Literature Validation’ below 10; ‘Topic Datasets’ score of zero or one.

Table 7: Patent Class Definitions

Number	Classification	Inventions Pertaining To:
(1)	Agriculture	The growth of crops and raising of livestock; fishing, forestry and hunting; horticulture; unspecified use of land
(2)	Apparel	Articles to be worn; articles of clothing for humans and animals; jewellery, broaches, and the like
(3)	Chemicals	The development of new chemicals, the applications of chemicals, or products developed by chemicals processes; organic and inorganic chemistry; gases; nuclear
(4)	Commodities	Consumable, durable, and non-durable goods which are not explicitly for industrial usage, with a focus on inventions to be sold in the market for private use; intangible services; recreational items
(5)	Construction	Building; tools for building; civil engineering; construction and building related accessories; building of infrastructure; construction of items of a physical nature
(6)	Electricity	The creation, management, and application of electricity; of electrical appliances, components, and instruments; aspects of electricity which do not overlap with other utilities; combinations of electricity with galvanism, magnetism and the like
(7)	Food	The production, treatment, and management of foodstuffs and beverages for consumption; tobacco
(8)	Hardware	Devices, objects, items or articles that provide a productivity-enhancing or labour-saving function requiring little or no manual interaction; Machine tools; Non-mechanical objects
(9)	Health	Improving the quality of life; life-saving medicines or apparatus; protection from ailments
(10)	Instruments	Measuring, gauging, weighing; general devices or objects which reduce the effort required to perform certain tasks; devices or objects which aid in productivity of labour; a tool or implement especially for precision work
(11)	Machinery	Machines which operate on mechanical power, and to their maintenance; processes conducted by machines
(12)	Manufacturing	The production of goods or items; large scale and small scale
(13)	Metal	Metallurgy; extracting metals from their ores; the application of chemical processes to metals, whether by producing, refining, galvanising or other such methods
(14)	Military	Weapons, armaments, armour, and other types of offensive or defensive articles
(15)	Mining	The construction of mines, their excavation, management, flood management, and extraction of natural resources; the raising and lowering of heavy bodies
(16)	Paper	The use of paper; methods which improve paper; the process of printing; paper and cardboard production, and to other such related items; physical record keeping; bills, cheques
(17)	Power	Generating, regulating, and applying energy for power, speed, or such related uses
(18)	Textiles	The creation of fabrics from processes of weaving, spinning, knitting, felting, etc, and their bleaching or dyeing, and treatment
(19)	Transportation	Facilitating speedy, or easier, travel across distances; transport infrastructure; packaging and storage of items for easier transport
(20)	Utility	The management of public systems, such as sewerage; the creation, management, and application of gas, heat, light, and water; the regulation of water, light, heat, gas, and electricity as public goods; and to inventions which encompass combinations of water, light, heat, gas, and electricity; fireproofing structures

Notes: Definitions are constructed using the word associations derived from the topic analysis methodology. Some classes could be further divided using these definitions, or further aggregated.

technical function. Inadequate descriptions lead to a subjective interpretation of how to apply our taxonomy. Such difficulties would deter adoption of the schema, and undermine results derived from its application. Topics represent word clusters tending to appear in combination with one another. Where a topic directly relates to a class the words comprising the topic act as descriptors. This uncovers the vocabulary used to connect a patent’s description with its intended classification, from which we derive our class definitions. Descriptions also grant the schema adaptability; classes can be further aggregated or disaggregated to aid future users apply an appropriate taxonomy that suits their research question. Table 7 presents our schema definitions.

5 Comparison of Taxonomies

We now examine whether the choice of patent taxonomy influences how we interpret patent statistics by examining all British patents granted from 1700 to 1850. We use this dataset for four reasons. First, the dataset has been used extensively within the historical innovation literature (Dutton, 1984; Sullivan, 1989, 1990; MacLeod, 2002; Nuvolari and Tartari, 2011; Meisenzahl and Mokyr, 2011; Bottomley, 2014a; Dowey, 2017; Khan, 2018). Second, prior studies have classified the patent data. Woodcroft (1860), *A Cradle of Inventions* (2009), and Nuvolari and Tartari (2011) have assigned alternative schemas to the data, allowing for comparisons with our own. Third, the dataset covers the traditional period of the Industrial Revolution (dated approximately 1760-1830). Any insights from this era are vital to our understanding of this phenomenon. Fourth, Britain’s patent system remains relatively unchanged over the period we analyse. Our results are therefore unlikely to be driven by any institutional changes.

To test for classification divergence, we need to be able to accurately replicate prior taxonomies. Any degree of inaccuracy severely hinders the conclusions that can be drawn from this exercise. To this end, we are capable of replicating the three aforementioned taxonomies: Nuvolari and Tartari (2011); *A Cradle of Inventions* (2009) – used in Dowey (2017); and Woodcroft (1860) – used in Sullivan (1989, 1990)

and Brunt et al. (2012). These existing taxonomies do not consistently classify patent data: the schemas have different classes and presumably different approaches to classifying patents (see the online appendix for a comparison of these taxonomies). Consequently, the alternative taxonomies may produce different results when analysing the same patents.

For the Nuvolari and Tartari (2011) schema, the authors kindly provided their classified patent data, allowing us to match their patent classes to our data. *A Cradle of Inventions (2009)* is a CD-ROM containing the entire population of British patents with their own taxonomy assigned, henceforth known as ‘COI’. Again, we were able to match up our schema to that from *A Cradle of Invention*. The Woodcroft (1860) schema is assigned based on a unique British patent identifier constructed by Woodcroft contained in our dataset. Finally, we apply the schema using our machine learning methodology. Topics were manually assigned to a class based on the alignment of their highest scoring words to the class descriptions. Patents were then assigned to classes based on their highest two scoring topics, denoted as “Topic-One” and “Topic-Two”. Thus, patents linked to a topic dominated by words such as ‘spin, cotton, fibrous, twist, wool’ would be classified as Textiles.¹⁰

To test for any potential divergences, we observe each schema in relation to six commonly examined patent characteristics concerning the nature of invention: the citations of patented inventions, the occupational status of patentees, the stock of patents held by patentees, the number of named inventors per patent, whether a patentee is considered an insider, and whether a patent is for a capital-saving invention – for brevity, we report results only for patent citations and capital-saving inventions, the remaining results are contained in the online appendix. Because each taxonomy does not have the exact same patent classes, we examine only those classes common to at least three schemas: Agriculture, Chemicals, Clothing (or Apparel), Engines (or Power), Food, Instruments, Medicines (or Health), Metal, Military, Mining, Paper, and Textiles.

¹⁰The online appendix details the alternative taxonomies and our own in more detail.

While we could focus on those classes common to all schemas, doing so would remove useful information concerning the existence and degree of classification divergence. For all ensuing regressions, we include all patent classes in each schema, but report only the common classes to clarify our analysis. In addition, all regressions are run on observations common to all patent schemas.

We contrast the regressions in three ways. First, we contrast results between taxonomies for a given patent characteristic: how does a common patent class coefficient's magnitude, sign, or significance change from one taxonomy to another? Second, we contrast whether the differences between taxonomies are the same across patent characteristics: if we observe a particular class's divergence for one patent characteristic, do we observe the same class divergence for other characteristics? Third, we contrast the differences between our Topic-One and Topic-Two classes. This highlights whether patents spillover technology boundaries: the greater the difference, the greater the spillover with other technologies is likely to be, suggesting it necessary to control for both topics simultaneously.

Our interpretation centres around comparisons of coefficient magnitude, statistical significance, and the direction of association across taxonomies for the same patent class. When examining coefficients, the size of any difference relative to the coefficients being compared will provide an indication of magnitude divergence. Changes in statistical significance across taxonomies highlights inconsistencies in the subject matter being classified. We consider this to be strong where coefficients fluctuate from no statistical significance to significance at the one per cent level. The direction of association will also help us understand the inconsistency of classification. Where coefficients all have the same sign, this highlights a degree of consistency, but where they differ it further emphasises the degree of classification inconsistency.

The taxonomies examined in this study are very similar, and so we would expect to observe no differences if patents are classified in the same manner. But, observing any divergence, no matter how small, suggests taxonomy similarity is strictly not enough to ensure consistency, and that divergences between less similar taxonomies are likely to be

considerably larger and perhaps more serious.¹¹

Unfortunately, we are unable to estimate the degree of divergence within the existing literature, nor exactly how existing results would change had our methodology been employed. Such an endeavour is beyond the scope of this study. The purpose of our comparative analysis is to identify whether classification divergence exists between a set of similar patent taxonomies. If it does, then wider divergence would also be expected between dissimilar taxonomies.

5.1 The Citations of Patented Inventions

First, we examine how the choice of taxonomy affects an analysis of the citations of patented inventions. In the innovation literature, patent citations are a popular proxy for patent quality or value (Hall et al., 2001, 2005; Lach and Schankerman, 2008; Bernstein, 2015; Kogan et al., 2017). In place of citations, the historical British literature has adopted the Woodcroft Reference Index (WRI), as pioneered by Nuvolari and Tartari (2011). This index lists how many contemporary scientific and trade journals referenced a particular patent within our dataset. The references are used to proxy for the technical and economic significance of a patented invention: more references signals a higher quality patent. Because the number of references artificially increases over time, we adopt the approach of Hall et al. (2005); Nuvolari and Tartari (2011), by weighting the total sum of references on a patent by the average number of references on all patents within a given time period; the time periods we use are those used in Nuvolari and Tartari (2011).¹² To ensure comparability, all regressions also use the same time periods as time controls.

The quality indicator is a count variable with a skewed distribution; many patents have few references, and few patents have many references. The negative binomial model accounts for skewness by relaxing the assumption that the mean and the variance are equal (Greene, 2008).¹³ Under this model, our dependent variable is the weighted

¹¹As a further examination of similarities between the alternative schemas, the online appendix reports a series of regression plots showing correlations between the alternative schemas for the patent characteristics we examine in this section.

¹²These time periods are as follows: 1700-1721; 1722-1741; 1742-1761; 1762-1781; 1782-1801; 1802-1811; 1812-1821; 1822-1831; 1832-1841; 1842-1850.

¹³We also test the relationship using the poisson model. The results from poisson are equivalent to

number of references for a given patent. Our control variables constitute: whether the patentee had a prior patent; the patentee's occupation; whether the patentee's occupation directly relates to the class of their invention; their nationality; and time controls. The explanatory variables are the classes associated with each schema. We represent patent classes with dummy variables, where Agriculture is the chosen baseline category.

Table 8 presents our results. Column 1 uses the Woodcroft schema; column 2 the NT schema; column 3 the COI taxonomy; column 4 the Topic-One taxonomy; column 5 the Topic-Two taxonomy; and column 6 controls for Topic-One and Topic-Two simultaneously (henceforth known as "CombinedTopics"). Future investigators who employ our schema and methodology should run three separate econometric specifications, using Topic-One, Topic-Two, and then both schemas together as a robustness check.

At first glance, the table posits consistencies for several classes. Metallurgy and Textiles patents are the most consistent across taxonomies, perhaps reflecting better defined technology and industry boundaries compared to other patent classes. Despite the consistency, both classes still display divergence. For Metallurgy, coefficient magnitude fluctuates considerably – Metal patents are in the range of 10 to 20 per cent more valuable than Agricultural ones, dependent on taxonomy. Similarly, under the COI schema only, Textiles patents are not statistically different from Agricultural ones.

After examining the remaining classes, we find that classification divergences exist. This divergence affects all aspects related to interpreting regression coefficients. The magnitude of coefficients fluctuates when comparing Mining inventions, for example. The COI schema suggests Mining patents have 20 per cent more references per patent compared to Agricultural patents. One reasonable interpretation is that capital-intensive inventions are of a greater quality.¹⁴ Topic-One, however, suggests Mining patents have nine per cent fewer references. Capital-intensive inventions, then, are of a lower quality

the negative binomial approach.

¹⁴Based on their titles, Mining patents were likely to be highly mechanised during this period. Such inventions are considered to be capital-intensive, as suggested by Khan (2005), because more capital than labour is required for their development.

Table 8: Negative Binomial: Dependent Variable is the Weighted Number of References per Patent

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	0.090*** (0.029)	0.105*** (0.023)	- -	0.017 (0.020)	0.071*** (0.020)	0.018 (0.012)
Clothing	-0.058 (0.038)	-0.098*** (0.034)	-0.046 (0.039)	-0.151*** (0.028)	0.033 (0.039)	-0.055* (0.032)
Engines	0.009 (0.052)	0.045 (0.040)	- -	0.002 (0.038)	0.038 (0.033)	0.010 (0.025)
Food	- -	0.021 (0.022)	0.026 (0.027)	0.115 (0.164)	0.134*** (0.051)	0.122 (0.087)
Instruments	- -	0.041* (0.023)	-0.014 (0.025)	-0.039** (0.019)	0.017 (0.023)	-0.036*** (0.012)
Medicines	-0.104*** (0.027)	-0.069*** (0.025)	-0.031 (0.029)	-0.067** (0.028)	-0.092 (0.057)	-0.098*** (0.028)
Metallurgy	0.186*** (0.030)	0.145*** (0.030)	- -	0.098*** (0.033)	0.107*** (0.038)	0.092*** (0.029)
Military	-0.051* (0.028)	-0.019 (0.033)	-0.034 (0.038)	-0.067* (0.035)	-0.010 (0.026)	-0.043** (0.019)
Mining	0.092 (0.103)	0.182*** (0.060)	0.202*** (0.056)	-0.090** (0.036)	0.004 (0.051)	-0.055* (0.030)
Paper	0.210*** (0.058)	0.077* (0.040)	0.060 (0.043)	0.023 (0.030)	0.033 (0.028)	0.021 (0.021)
Textiles	-0.111*** (0.042)	-0.066* (0.039)	-0.039 (0.031)	-0.071*** (0.021)	-0.058*** (0.022)	-0.071*** (0.013)
Constant	-0.031 (0.071)	-0.101 (0.073)	-0.084 (0.065)	-0.041 (0.059)	-0.058 (0.052)	-0.034 (0.052)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	12,937	12,937	12,937	12,937	12,937	12,937
Pseudo R-Squared	0.00667	0.00371	0.00287	0.00292	0.00271	0.00325

Notes: The table shows how the quality of patented inventions varies by technology group. The dependent variable is the weighted number of references per 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 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.

compared to Agricultural ones.

Statistical significance also exhibits inconsistency. Chemicals patents, for example, show a statistically significant association under the NT, Topic-Two and Woodcroft schemas. For Topic-One and CombinedTopics, however, Chemicals patents are not statistically distinguishable from Agricultural patents, in terms of their respective number of references. Such a result could lead investigators to consider Chemical patents as being no different from Agricultural patents. Similar observations are found when interpreting Clothing, Food, Instruments, Medicines, Military, Paper, and Textile related patents.

The direction of association of coefficients is also subject to divergence. Three classes exhibit inconsistency regarding the direction of association: Clothing, Instruments, and Mining. Mining and Instruments are perhaps the most divergent classes, as not only does their direction of association fluctuate across taxonomies, but so too does their statistical significance – both positive and negative statistically significant associations are shown. Depending on which schema had been adopted, investigators could find completely contradictory results.

In addition, the results suggest the need for two classes. Should patents have the same assigned class for both Topics then we would expect to observe similar results. However, contrasting Topic-One and Topic-Two shows that, for the majority of classes, there is a significant difference between the coefficients. This suggests that most patents within the affected patent classes have characteristics that spillover into other technology boundaries. In such cases, it is necessary to control for a second class to capture potential spillovers. The potential for spillovers can also be observed in the topic weightings for each patent. A single dominant topic, or multiple dominant topics mapping to the same class, suggest a single classification is sufficient. However, in other cases making a single classification risks ignoring this spillover potential.

5.2 Capital Saving Patents

We next examine whether a patent is explicitly for a capital-saving invention. Invention during the Industrial Revolution has been viewed as a response to relative factor prices. According to Allen (2009), Britain had relatively high labour costs and cheap capital. The relative factor prices, Allen proposes, led to a rise in labour-saving invention and thus caused the Industrial Revolution.¹⁵ This argument has been contested by Mokyr (2009), who points out that a significant number of inventions were not labour-saving. In her seminal work, MacLeod (2002) notes that the vast majority of patented inventions granted in Britain prior to 1800 were for capital-saving inventions.

Understanding which types of technologies are likely to benefit from capital-saving invention helps us understand the nature of invention. Do people choose to invent, and then also patent, because certain productive activities have expensive capital inputs? Considering the rise in capital-related invention during the Industrial Revolution, examining patent records is an important means to observe this type of inventive behaviour. Consequently, understanding how interpretations from such data can be directed by the choice of taxonomy is equally important.

Following the methodology of MacLeod (2002), we construct a dummy variable for patents which explicitly reference a form of capital-saving embodied in their invention.¹⁶ A value of zero represents patents which did not have a stated capital-saving aim, while labour-saving inventions are controlled for in the analysis.¹⁷

We use a probit regression model to examine the probability of patent classes being associated with capital-saving inventions. 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.

¹⁵Labour-saving inventions are those which intend to reduce the amount of labour involved in the production process. Conversely, capital-saving inventions are intended to cheapen the cost of using capital.

¹⁶MacLeod constructed a list of stated aims derived from the titles of patents granted prior to 1800. This includes: saving time; saving fuel; saving on raw materials; increasing output; increasing power; reliability of equipment; regularity of output; saving on running costs in general; and saving on fixed capital. We follow this approach and only classify a patent as capital-saving if it states any of the aforementioned aims in its title.

¹⁷We do not analyse labour-saving patents because they are too few.

The results, reported in Table 9, exhibit classification divergence. When examining capital-saving patents, divergence is significantly more extreme than prior results (see also characteristics examined in the online appendix). This is likely because we have fewer available observations: capital-saving statements cease post-1830, and so drop from our analysis. This may inhibit the comparability of divergence with the other taxonomies, but we are still able to contrast across schemas for this metric.

Coefficient magnitude is subject to divergence. For some classes, such as Metal, Mining, Paper, or Textiles patents, this divergence seems relatively small, while other classes, such as Chemicals, Clothing, and Medicines, exhibit a greater degree of divergence. Observing Medicines, for example, the Topic-Two schema suggests that Medicinal patents are on average three per cent less likely to be capital-saving than Agricultural patents. However, under the Woodcroft schema, the coefficient estimate for Medicinal patents is instead 29 per cent – a considerable difference in magnitude.

Statistical significance fluctuates across taxonomies. For most of the common classes, no statistical significance and statistical significance at the one per cent level are reported. For example, Clothing patents are statistically significant at either the one per cent or five per cent level under the Woodcroft, NT, COI, and CombinedTopics schemas. For Topic-Two the association is not statistically significant at any conventional level. Only Textile patents remain consistent, exhibiting no statistically significant associations across any of the alternative taxonomies.

The direction of association is also inconsistent across taxonomies. For Chemicals, Clothing, Food, Mining, and Paper, both positive and negative associations are reported. In several of these classes there is an almost even divide between the direction of association. For example, the Woodcroft and NT schemas report a positive association between Chemical patents and capital-saving aims. By contrast, the Topic-One, Topic-Two, and CombinedTopics schemas report the opposite. Additionally, differences between Topic-One and Topic-Two coefficients continue to suggest many patents have characteristics which spillover multiple technology boundaries, further supporting the need to include an additional secondary classification in patent analyses.

Table 9: Probit: Dependent Variable is a Dummy representing a Capital-saving Patent

VARIABLES	(1) Woodcroft	(2) NT	(3) COI	(4) Topic-One	(5) Topic-Two	(6) CombinedTopics
Chemicals	0.099* (0.056)	0.093*** (0.031)	- -	-0.055** (0.027)	-0.127*** (0.036)	-0.165*** (0.032)
Clothing	-0.191*** (0.028)	-0.073** (0.032)	-0.102*** (0.026)	0.059** (0.028)	0.050 (0.032)	0.062*** (0.019)
Engines	0.021 (0.048)	0.063** (0.026)	- -	0.077** (0.038)	0.004 (0.033)	0.039*** (0.014)
Food	- -	0.071 (0.049)	-0.025 (0.028)	0.286* (0.152)	-0.001 (0.067)	0.106** (0.051)
Instruments	- -	0.066* (0.038)	0.024 (0.042)	0.034 (0.029)	-0.001 (0.028)	0.021 (0.015)
Medicines	-0.289*** (0.033)	-0.087*** (0.033)	-0.143*** (0.028)	-0.092*** (0.017)	-0.035 (0.042)	-0.076*** (0.029)
Metal	0.059 (0.036)	0.071* (0.038)	- -	0.081** (0.032)	0.036 (0.034)	0.067*** (0.021)
Military	0.061* (0.036)	0.034 (0.030)	0.006 (0.033)	0.009 (0.029)	0.032 (0.041)	0.011 (0.027)
Mining	0.072 (0.071)	0.118 (0.118)	-0.055 (0.054)	0.088** (0.039)	0.076 (0.052)	0.069*** (0.017)
Paper	0.013 (0.068)	-0.020 (0.033)	-0.070** (0.031)	-0.032 (0.025)	-0.015 (0.040)	-0.032 (0.020)
Textiles	-0.009 (0.038)	-0.013 (0.028)	-0.024 (0.028)	-0.011 (0.026)	-0.030 (0.033)	-0.010 (0.019)
Time	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	5,379	5,379	5,379	5,379	5,379	5,379
Pseudo R-Squared	0.158	0.119	0.109	0.103	0.103	0.110

Notes: The table shows whether a capital-saving patent varies by technology group. The dependent variable is a dummy variable indicating whether a patent explicitly stated its purpose was to save capital, following the approach of MacLeod (2002). There are fewer observations in this instance because capital or labour saving statements cease post-1830 and so are dropped from the analysis. 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 1700-1830.

6 Discussion

To our knowledge, the present study is the first to show classification divergence exists in the field of innovation studies. Prior studies have not explicitly addressed the potential consequences of their choice of taxonomy. This is a serious concern, as statistical significance, direction of influence, and coefficient magnitude are sensitive to the chosen schema. In addition, classification divergence is not consistent: the degree of divergence can vary dependent on what patent characteristics are being investigated. Depending on which schemas are used, and which patent characteristics are examined, investigators could find conflicting results.

The extent of classification divergence within the literature is uncertain. Without a complete understanding of how authors construct their taxonomies, for what purpose, and the method they use for classifying patents, we cannot determine how significant the divergence might be. Most academic studies do not provide such detail. Therefore, prior research articles which do not expressly describe their taxonomy should be interpreted with caution. Our findings have implications for the field of innovation history, as well as economic policymaking more generally.

Innovation historians examine patent data to better understand inventive behaviour and the nature of invention, typically during periods of rapid technological change. Patents are one of the few available sources of time-series invention data and, as such, are popularly used throughout innovation history literature. This gives rise to a significant number of alternative taxonomies which cannot be accurately replicated. Given the results observed in this study, the number of alternative schemas serves to complicate our interpretation of patent studies.

The existence of classification divergence within the innovation history literature is concerning because it calls into question our collective understanding of patent systems as a means to foster innovation. Given the recent re-evaluation of the importance of Britain's patent system in encouraging the Industrial Revolution (Bottomley, 2014b, 2019), and the increasing use of patent indicators to measure inventive behaviour and technological change more generally (Khan, 2013a; Nuvolari and Vasta, 2015; Dowey,

2017; Murfitt, 2017; Donges and Selgert, 2019; Lane, 2019, for example), the implications of classification divergence are significant.

To highlight the severity of classification divergence, we can use our results to contrast how our understanding of the nature of invention may be subject to the choice of schema. Considering the Industrial Revolution represents a shift away from agricultural economies toward mechanised ones, we compare Mining patents against Agricultural patents. Mining was an important industry during the British Industrial Revolution, and responsible for a number of important technological advancements (Nuvolari, 2004; Allen, 2009; Mokyr, 2009; de Pleijt et al., 2019). Indeed, James Watt's famous separate condenser significantly improved the efficiency of steam engines: one of the key technologies of the Industrial Revolution, and a mining invention.

Considering first the results from the Woodcroft schema, they reasonably suggest that: Mining patents were at least of the same economic value as Agricultural patents (Table 8) and Mining patents were not any more likely to be capital-saving (Table 9). By contrast, the Topic-One schema reasonably suggests that Mining patents were less economically valuable than Agricultural patents (Table 8) and more likely to be capital-saving (Table 9).¹⁸

The two schemas present competing results, which lead to competing interpretations concerning the Industrial Revolution and the importance of Mining innovation. For the economic historian, the competing results can make it difficult to ascertain the nature of invention, and by extension, the reasons behind British inventive activity.

Aside from the implications for innovation historians, we argue our results have implications for current innovation studies and economic policymaking. To our knowledge, there is no commonly accepted patent taxonomy used throughout the innovation literature, and so may also be subject to classification divergence.

Innovation scholars are interested in understanding the process of innovation and how best to encourage it. Although patents are not the only metric available that captures innovative activity, they remain a popular one (Griliches, 1990). From a survey of US

¹⁸The online appendices contain further results concerning the nature of invention.

manufacturing firms, Cohen et al. (2000) find that patents are more useful for possible rent-seeking motives, such as preventing competitors from patenting related inventions, or to force competitors into negotiations with the patent holders, and different strategies are used in relation to different technologies. Another possible strategy is to use patents solely for the purpose of litigating or to defend against possible patent litigation from competitors. Lanjouw and Schankerman (2001) find litigation rates vary across patent technology groups, while more recent evidence suggests certain firms purchase patents solely for purposes of enforcing them against infringers, thereby appropriating their profits (Fischer and Henkel, 2012; Galasso and Schankerman, 2015). Such strategic patenting has become an important explanation for the modern day patenting behaviour of firms (Kingston, 2001; Arundel and Patel, 2003; Blind et al., 2006; Boldrin and Levine, 2008).

The decision to acquire a patent is therefore related to the technology of the invention in question. Indeed, patenting incentives in the modern innovation literature are likely more numerous and complex than those identified in innovation history. Such complexity makes accurately identifying relevant technology groups in a consistent manner important. To understand and encourage innovative behaviour, scholars and policymakers need to understand the nature of innovation and what role patents play within it. If scholars do not classify their patent data in a consistent or transparent manner, then the comparability of existing studies may be called into question.

Policymakers too, may be confounded by such issues. Returning to our earlier example, a policymaker tasked with designing measures to encourage innovation would draw wholly different conclusions from research of this historical period based on their choice of taxonomy. Of course we are not suggesting that such decisions should be made on the basis of single sources of evidence, or single episodes from history. However, it does highlight the potential implications classification divergence has for prescribing policy, which may result in the misdirection or suboptimal allocation of important resources, or in extreme cases, inadvertently hinder rather than encourage innovation.

Our recommendations for future patent investigations are as follows. Firstly, creators of new taxonomies should describe how they design them, ensuring potential

divergences can be identified and their methods are replicable. Descriptions need to accompany patent classes to ensure a consistent classification of patent data throughout the literature. Secondly, mitigating potential divergences requires adopting a universal schema. While the classification procedure may also depend on the research question being investigated, there still needs to be a consistent classification of the same technologies. The taxonomy produced here is a useful starting point, and adaptable for future studies. Thirdly, subjectivity can be reduced by employing machine learning techniques to improve the consistency of patent classification. Finally, topic analysis provides a means to both identify appropriate classes and omitted classes, and to perform the classification of patent datasets in useful ways for economic analysis of innovation.

7 Conclusion

Our goal has been: to document methods of taxonomy construction; to design and develop a new and adaptable time-invariant patent taxonomy in a clear and transparent manner; to develop a new method for classifying all patent data consistently; and to show that classification divergence exists. We recommend our methodology and taxonomy be used in future studies. We acknowledge, however, that our schema may not be applicable to every study. In such cases, future investigators should describe either how they adapt our schema or any new taxonomies they produce. The machine learning techniques herein described are adaptable and adoptable for any future researchers.

The implications of classification divergence are likely to be profound for the patents and innovation literature. Classification divergence exists, at least, in the long-run British patent data studied here. Whether divergence exists in other datasets necessitates a re-examination of the existing literature, for clarification. In the case where divergence is small, interpreting the literature is less problematic, and deriving appropriate policy measures would remain possible. However, in the extreme case, where all studies are subject to divergence, the external validity of studies must be questioned. Given the

results shown here, the extreme case seems more likely.

When studies are not comparable, appropriate policy measures cannot be readily prepared. We recommend existing studies, where possible, be re-examined using our schema and methodology. This is not to say our schema is “right”, as there can be no objective measure of this. Our schema is transparent, however, making it straightforward for any subsequent studies to make use of it, or draw from it, as they see fit, while our methodology is consistent and replicable. Human error and human effort is substantially minimised using our machine learning approach. Related patents will always be identified and grouped together. Compared to humans, the machine is more consistent.

References

- A Cradle of Inventions: British Patents from 1617 to 1894 (2009), *Stevenage, UK: Metal Finishing Information Services Ltd.*
- Akcigit, U., J. Grigsby, and T. Nicholas (2017), ‘The rise of American ingenuity: Innovation and inventors of the Golden Age,’ *National Bureau of Economic Research Working Paper Series No. 23047.*
- Allen, R. C. (2009), *The British Industrial Revolution in global perspective.* Cambridge: Cambridge University Press.
- Arts, S., B. Cassiman, and J. C. Gomez (2018), ‘Text matching to measure patent similarity,’ *Strategic Management Journal*, **39**(1), 62–84.
- Arts, S. and R. Veugelers (2015), ‘Technology familiarity, recombinant novelty, and breakthrough invention,’ *Industrial and Corporate Change*, **24**(6), 1215–1246.
- Arundel, A. and P. Patel (2003), ‘Strategic patenting,’ In *Background report for the “Trend Chart Policy Benchmarking Workshop” New Trends in IPR Policy.*
- Bailey, M. F. (1946), ‘History of classification of patents,’ *Journal of the Patent Office Society*, **28**(7), 463–507.

- Balsmeier, B., M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G.-C. Li, S. Lück, D. O'Reagan, B. Yeh, and Others (2018), 'Machine learning and natural language processing on the patent corpus: Data, tools, and new measures,' *Journal of Economics & Management Strategy*, **27**(3), 535–553.
- Basberg, B. L. (1997), 'Creating a patent system in the European periphery: The case of Norway, 1839-1860,' *Scandinavian Economic History Review*, **45**(2), 142–158.
- Basberg, B. L. (2006), 'Patenting and early industrialization in Norway, 1860-1914. Was there a linkage?,' *Scandinavian Economic History Review*, **54**(1), 4–21.
- Baten, J., A. Spadavecchia, J. Streb, and S. Yin (2007), 'What made southwest German firms innovative around 1900? Assessing the importance of intra- and inter-industry externalities,' *Oxford Economic Papers*, **59**(suppl 1), i105–i126.
- Benner, M. and J. Waldfogel (2008), 'Close to you? Bias and precision in patent-based measures of technological proximity,' *Research Policy*, **37**(9), 1556–1567.
- Bernstein, S. (2015), 'Does going public affect innovation?,' *The Journal of Finance*, **70**(4), 1365–1403.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), 'Latent dirichlet allocation,' *Journal of machine Learning research*, **3**, 993–1022.
- Blind, K., J. Edler, R. Frietsch, and U. Schmoch (2006), 'Motives to patent: Empirical evidence from Germany,' *Research Policy*, **35**(5), 655–672.
- Blit, J. and M. Packalen (2019), 'A machine learning analysis of the geographic localization of knowledge flows,' *Northwestern University Working Paper*.
- Boldrin, M. and D. K. Levine (2008), *Against intellectual monopoly*. New York: Cambridge University Press.
- Boschma, R., P.-A. Balland, and D. F. Kogler (2014), 'Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010,' *Industrial and Corporate Change*, **24**(1), 223–250.

- Bottomley, S. (2014a), 'Patenting in England, Scotland and Ireland during the Industrial Revolution, 1700-1852,' *Explorations in Economic History*, **54**, 48–63.
- Bottomley, S. (2014b), *The British patent system during the Industrial Revolution 1700-1852: From privilege to property*. Cambridge University Press.
- Bottomley, S. (2019), 'The returns to invention during the British Industrial Revolution,' *Economic History Review*, **72**(2), 510–530.
- Brunt, L., J. Lerner, and T. Nicholas (2012), 'Inducement prizes and innovation,' *Journal of Industrial Economics*, **60**(4), 657–696.
- Burhop, C. and N. Wolf (2013), 'The German market for patents during the "Second Industrialization", 1884-1913: A gravity approach,' *Business History Review*, **87**(1), 69–93.
- Cantwell, J. (2000), 'Technological lock-in of large firms since the interwar period,' *European Review of Economic History*, **4**(2), 147–174.
- 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, H., G. Zhang, D. Zhu, and J. Lu (2017), 'Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014,' *Technological Forecasting and Social Change*, **119**, 39–52.
- 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.
- Clark, G. (2007), *A farewell to alms: A brief economic history of the world*. Princeton University Press.
- Cohen, W. M., R. R. Nelson, and J. P. Walsh (2000), 'Protecting their intellectual assets:

- Appropriability conditions and why US manufacturing firms patent (or not),’ *National Bureau of Economic Research No. 7552*.
- Comino, S., A. Galasso, and C. Graziano (2017), ‘The diffusion of new institutions: Evidence from Renaissance Venice’s patent system,’ *National Bureau of Economic Research Working Paper Series No. 24118*.
- Costantini, V., F. Crespi, and Y. Curci (2015), ‘A keyword selection method for mapping technological knowledge in specific sectors through patent data: the case of biofuels sector,’ *Economics of Innovation and New Technology*, **24**(4), 282–308.
- Davids, K. (2000), ‘Patents and patentees in the Dutch republic between c.1580 and 1720,’ *History and Technology*, **16**(3), 263–283.
- de Pleijt, A., A. Nuvolari, and J. Weisdorf (2019), ‘Human capital formation during the First Industrial Revolution: Evidence from the use of steam engines,’ *Journal of the European Economic Association*, **17**, 1–61.
- Donges, A. and F. Selgert (2019), ‘Technology transfer via foreign patents in Germany, 1843-1877,’ *The Economic History Review*, **72**(1), 182–208.
- Dowey, J. (2017), *Mind over matter: Access to knowledge and the British Industrial Revolution*. Ph D thesis, The London School of Economics and Political Science.
- Dutton, H. I. (1984), *The patent system and inventive activity during the Industrial Revolution, 1750-1852*. Manchester University Press.
- EPO (2017), *Patents and the fourth Industrial Revolution*. Munich. Available from [http://documents.epo.org/projects/babylon/eponet.nsf/0/17FDB5538E87B4B9C12581EF0045762F/\\$File/fourth_industrial_revolution_2017__en.pdf](http://documents.epo.org/projects/babylon/eponet.nsf/0/17FDB5538E87B4B9C12581EF0045762F/$File/fourth_industrial_revolution_2017__en.pdf)
- Feng, F. S. (2019), ‘The proximity of ideas: an analysis of patent text using machine learning,’ *NYU Stern Working Paper*.

- Fischer, T. and J. Henkel (2012), 'Patent trolls on markets for technology – An empirical analysis of NPEs' patent acquisitions,' *Research Policy*, **41**(9), 1519–1533.
- Frederico, J. P. (1964), 'Historical patent statistics,' *Journal of the Patent Office Society*, **46**(3), 89–172.
- Galasso, A. and M. Schankerman (2015), 'Patents and cumulative innovation: Causal evidence from the courts,' *The Quarterly Journal of Economics*, **130**(1), 317–369.
- Greasley, D. and L. Oxley (2010), 'Knowledge, natural resource abundance and economic development: Lessons from New Zealand 1861-1939,' *Explorations in Economic History*, **47**(4), 443–459.
- Greene, W. (2008), 'Functional forms for the negative binomial model for count data,' *Economics Letters*, **99**(3), 585–590.
- Griliches, Z. (1990), 'Patent statistics as economic indicators: A survey,' *Journal of Economic Literature*, **28**(4), 1661–1707.
- Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001), 'The NBER patent citation data file: Lessons, insights and methodological tools,' *National Bureau of Economic Research Working Paper Series No. 8498*.
- Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2005), 'Market value and patent citations,' *The RAND Journal of Economics*, **36**(1), 16–38.
- 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.
- Johnson, D. K. N. (2002), 'The OECD Technology Concordance (OTC): Patents by industry of manufacture and sector of use,' *OECD Science, Technology and Industry Working Papers No. 2002/0*.

- 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.
- Kaplan, S. and K. Vakili (2012), ‘Identifying breakthroughs: using topic modeling
to distinguish the cognitive from the economic,’ *Academy of Management
Proceedings*, **2012**, 1–1.
- Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018), ‘Measuring technological
innovation over the long run,’ *National Bureau of Economic Research Working Paper
Series No. 25266*.
- Khan, B. Z. (2000), ‘ “Not for Ornament” : Patenting activity by nineteenth-century
women inventors,’ *The Journal of Interdisciplinary History*, **31**(2), 159–195.
- Khan, B. Z. (2005), *The democratization of invention: Patents and copyrights in
American economic development, 1790-1920*. Cambridge University Press.
- Khan, B. Z. (2013a), ‘Going for gold. Industrial fairs and innovation in the nineteenth-
century United States,’ *Revue économique*, **64**(1), 89–113.
- Khan, B. Z. (2013b), ‘Selling ideas: An international perspective on patenting and
markets for technological innovations, 1790-1930,’ *Business History Review*, **87**(1),
39–68.
- Khan, B. Z. (2014), ‘Inventing in the shadow of the patent system: Evidence from
19th-Century patents and prizes for technological innovations,’ *National Bureau of
Economic Research Working Paper Series No. 20731*.
- Khan, B. Z. (2015a), ‘Inventing prizes: A historical perspective on innovation awards
and technology policy,’ *Business History Review*, **89**(4), 631–660.
- 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. (2016), 'Invisible women: Entrepreneurship, innovation, and family firms in nineteenth-century France,' *The Journal of Economic History*, **76**(1), 163–195.
- Khan, B. Z. (2017), 'Prestige and profit: The Royal Society of Arts and incentives for innovation, 1750-1850,' *National Bureau of Economic Research Working Paper Series No. 23042*.
- 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*.
- Khimich, N. V. and R. Bekkerman (2016), 'Technological similarity and stock return cross-predictability: evidence from patents' big data,' *Available at SSRN 3147218*.
- Kingston, W. (2001), 'Innovation needs patents reform,' *Research Policy*, **30**(3), 403–423.
- 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.
- Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017), 'Technological innovation, resource allocation, and growth,' *The Quarterly Journal of Economics*, **132**(2), 665–712.
- Kortum, S. and J. Putnam (1997), 'Assigning patents to industries: Tests of the Yale Technology Concordance,' *Economic Systems Research*, **9**(2), 161–176.
- Lach, S. and M. Schankerman (2008), 'Incentives and invention in Universities,' *The RAND Journal of Economics*, **39**(2), 403–433.

- Lane, J. (2019), ‘Secrets for sale? Innovation and the nature of knowledge in an early industrial district: The potteries, 1750-1851,’ *Enterprise & Society*, 1–46.
- Lanjouw, J. O. and M. Schankerman (2001), ‘Characteristics of patent litigation: A window on competition,’ *The RAND Journal of Economics*, **32**(1), 129–151.
- Lehmann-Hasemeyer, S. and J. Streb (2016), ‘The Berlin Stock Exchange in Imperial Germany: A market for new technology?,’ *American Economic Review*, **106**(11), 3558–3576.
- Leydesdorff, L., D. F. Kogler, and B. Yan (2017), ‘Mapping patent classifications: portfolio and statistical analysis, and the comparison of strengths and weaknesses,’ *Scientometrics*, **112**(3), 1573–1591.
- Lybbert, T. J. and N. J. Zolas (2014), ‘Getting patents and economic data to speak to each other: An ‘Algorithmic Links with Probabilities’ approach for joint analyses of patenting and economic activity,’ *Research Policy*, **43**(3), 530–542.
- MacLeod, C. (2002), *Inventing the Industrial Revolution: The English patent system, 1660-1800*. Cambridge University Press.
- Magee, G. B. (1999), ‘Technological development and foreign patenting: Evidence from 19th-century Australia,’ *Explorations in Economic History*, **36**(4), 344–359.
- Magerman, T., B. V. Looy, and K. Debackere (2015), ‘Does involvement in patenting jeopardize one’s academic footprint? An analysis of patent-paper pairs in biotechnology,’ *Research Policy*, **44**(9), 1702–1713.
- Marco, A. C., M. Carley, S. Jackson, and A. Myers (2015), ‘The USPTO historical patent data files: Two centuries of innovation,’ *U.S. Patent and Trademark Office Working Paper 2015-1*.
- McNamee, R. C. (2013), ‘Can’t see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example,’ *Research Policy*, **42**(4), 855–873.

- Meisenzahl, R. and J. Mokyr (2011), ‘The rate and direction of invention in the British Industrial Revolution: Incentives and institutions,’ *National Bureau of Economic Research Working Paper Series No. 16993*.
- 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. (2005), ‘How do patent laws influence innovation? Evidence from nineteenth-century World’s Fairs,’ *The American Economic Review*, **95**(4), 1214–1236.
- Moser, P. (2012), ‘Innovation without patents: Evidence from World’s Fairs,’ *The Journal of Law & Economics*, **55**(1), 43–74.
- Murfitt, S. E. (2017), *The English patent system and early railway technology 1800-1852*. PhD thesis, University of York.
- Nanda, R. and T. Nicholas (2014), ‘Did bank distress stifle innovation during the Great Depression?,’ *Journal of Financial Economics*, **114**(2), 273–292.
- Nicholas, T. (2008), ‘Does innovation cause stock market runups? Evidence from the Great Crash,’ *The American Economic Review*, **98**(4), 1370–1396.
- 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. (2011a), ‘Did R&D firms used to patent? Evidence from the first innovation surveys,’ *Journal of Economic History*, **71**(4), 1032–1059.
- Nicholas, T. (2011b), ‘Independent invention during the rise of the corporate economy in Britain and Japan,’ *Economic History Review*, **64**(3), 995–1023.

- Nicholas, T. (2011c), ‘The origins of Japanese technological modernization,’ *Explorations in Economic History*, **48**(2), 272–291.
- Nuvolari, A. (2004), ‘Collective invention during the British Industrial Revolution: The case of the Cornish pumping engine,’ *Cambridge Journal of Economics*, **28**(3), 347–363.
- Nuvolari, A. and V. Tartari (2011), ‘Bennet Woodcroft and the value of English patents, 1617-1841,’ *Explorations in Economic History*, **48**(1), 97–115.
- Nuvolari, A. and M. Vasta (2015), ‘Independent invention in Italy during the Liberal Age, 1861-1913,’ *Economic History Review*, **68**(3), 858–886.
- 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.
- Packalen, M. and J. Bhattacharya (2012), ‘Words in patents: Research inputs and the value of innovativeness in invention,’ *National Bureau of Economic Research Working Paper Series No. 18494*.
- Pearce, E. (1957), *History of the Standard Industrial Classification*. Technical report, Executive Office of the President. Office of Statistical Standards, Washington.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), ‘Scikit-learn: Machine learning in Python,’ *Journal of Machine Learning Research*, **12**, 2825–2830.
- Rigby, D. L. (2015), ‘Technological relatedness and knowledge space: Entry and exit of US cities from patent classes,’ *Regional Studies*, **49**(11), 1922–1937.

- Righi, C. and T. Simcoe (2019), ‘Patent examiner specialization,’ *Research Policy*, **48**(1), 137–148.
- Ruckman, K. and I. McCarthy (2016), ‘Why do some patents get licensed while others do not?,’ *Industrial and Corporate Change*, **26**(4), 667–688.
- Sáiz, P. (2014), ‘Did patents of introduction encourage technology transfer? Long-term evidence from the Spanish innovation system,’ *Cliometrica*, **8**(1), 49–78.
- Schmoch, U. (2008), Concept of a technology classification for country comparisons. *Final report to the World Intellectual Property Organisation*.
- Schmoch, U., F. Laville, P. Patel, and R. Frietsch (2003), Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*.
- Schmookler, J. (1966), *Invention and economic growth*. Harvard University Press.
- Scotchmer, S. (1991), ‘Standing on the shoulders of giants: Cumulative Research and the patent law,’ *The Journal of Economic Perspectives*, **5**(1), 29–41.
- Scotchmer, S. (2004), *Innovation and incentives*. MIT Press.
- Sokoloff, K. L. (1988), ‘Inventive activity in early industrial America: Evidence from patent records, 1790-1846,’ *The Journal of Economic History*, **48**(4), 813–850.
- 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.
- Streb, J., J. Baten, and S. Yin (2006), ‘Technological and geographical knowledge spillover in the German empire 1877-1918,’ *Economic History Review*, **59**(2), 347–373.
- Strumsky, D., J. Lobo, and S. van der Leeuw (2012), ‘Using patent technology codes to study technological change,’ *Economics of Innovation and New Technology*, **21**(3), 267–286.
- Sullivan, R. J. (1989), ‘England’s ‘Age of Invention’: The acceleration of patents and

- patentable invention during the Industrial Revolution,' *Explorations in Economic History*, **26**(4), 424–452.
- Sullivan, R. J. (1990), 'The revolution of ideas: Widespread patenting and invention during the English Industrial Revolution,' *The Journal of Economic History*, **50**(2), 349–362.
- Suominen, A., H. Toivanen, and M. Seppänen (2017), 'Firms' knowledge profiles: Mapping patent data with unsupervised learning,' *Technological Forecasting and Social Change*, **115**, 131–142.
- Thompson, P. and M. Fox-Kean (2005), 'Patent citations and the geography of knowledge spillovers: A reassessment: Reply,' *The American Economic Review*, **95**(1), 465–466.
- Van Leeuwen, M. H. and I. Maas (2011), *HISCLASS: A historical international social class scheme*. Leuven University Press.
- Venugopalan, S. and V. Rai (2015), 'Topic based classification and pattern identification in patents,' *Technological Forecasting and Social Change*, **94**, 236–250.
- Verspagen, B., T. Van Moergastel, and M. Slabbers (1994), 'MERIT concordance table: IPC-ISIC (rev. 2),' *MERIT Research Memorandum February*.
- WIPO (1992). The International Patent Classification (IPC). *Journal of the Patent and Trademark Office Society* **74**(7), 481–483.
- WIPO (2016), *Guide to the International Patent Classification*. Available from https://www.wipo.int/edocs/pubdocs/en/wipo_guide_ipc_2016.pdf
- 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.
- Wu, C.-H., Y. Ken, and T. Huang (2010), 'Patent classification system using a new hybrid genetic algorithm support vector machine,' *Applied Soft Computing*, **10**(4), 1164–1177.
- Wu, J.-L., P.-C. Chang, C.-C. Tsao, and C.-Y. Fan (2016), 'A patent quality analysis and

classification system using self-organizing maps with support vector machine,' *Applied Soft Computing*, **41**, 305–316.

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

Youn, H., D. Strumsky, L. M. A. Bettencourt, and J. Lobo (2015), 'Invention as a combinatorial process: Evidence from US patents,' *Journal of The Royal Society Interface*, **12**(106), 20150272.

Younge, K. A. and J. M. Kuhn (2016), 'Patent-to-patent similarity: A vector space model,' *SSRN Working Paper No. 270923*.

Acknowledgements

Billington thanks Chris Colvin and Christopher Coyle for their detailed feedback and guidance in developing this paper, and Alessandro Nuvolari for hosting him on a visit to the Sant'Anna School of Advanced Studies, Pisa, in March 2018. Thanks to: a senior patent examiner for assisting in the construction of our patent classification schema, Owen Sims for his assistance with database software, Gerben Bakker, Graham Brownlow, Alan de Bromhead, Nola Hewitt-Dundas, David Jordan, and John Turner for their helpful comments. Thanks also to seminar participants at Queen's University Belfast (June, 2017). Finally, the authors would like to acknowledge the valuable feedback received from the anonymous reviewers. The authors will make their code available upon request.

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.¹⁹ 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:

¹⁹For example, ‘cultivate’ and ‘cultivating’ have the same base ‘cultivat’ but different stems, and would be observed as unique words without stemming.

NMF: steam, engine, rotary, rotatory, marine, navigable, condense, power, vapour, 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.²⁰ 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.

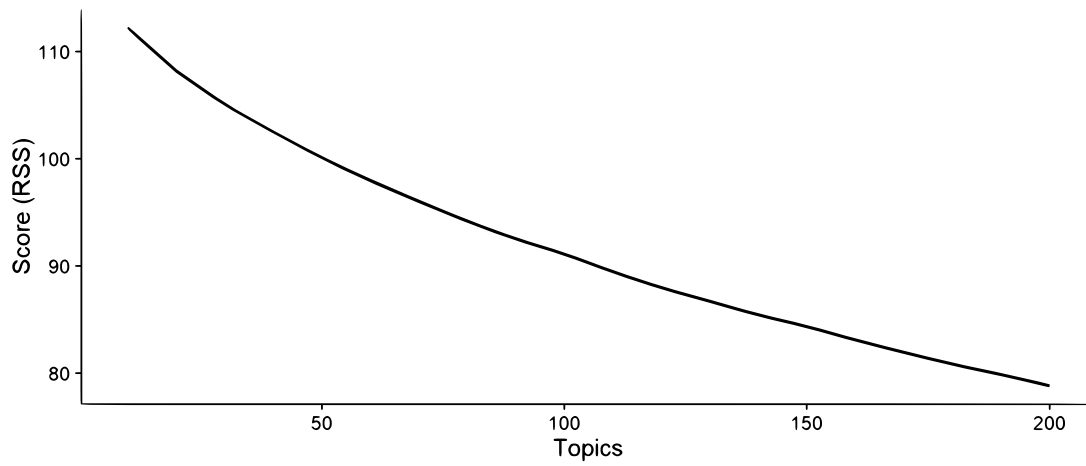
The RSS measures the quality of the approximation to the original document term frequency matrix, where a higher score suggests a less accurate representation. This metric decreases with each additional topic. In the case where there is a hidden number of groups, we may observe an improvement in the 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:

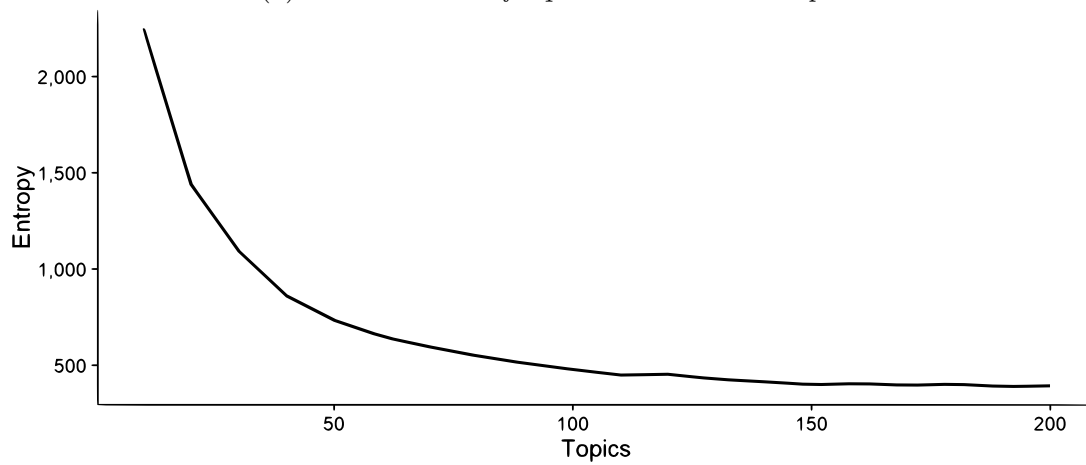
$$H(M) = \sum_{i=1}^t -P(i)\log P(i) \quad (2)$$

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

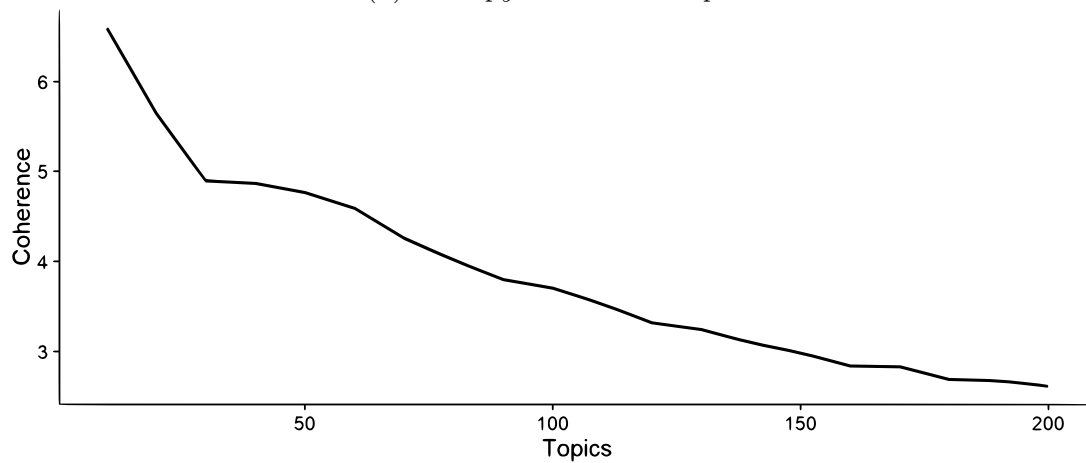
²⁰We generate the optimal number of topics from our British dataset, described in Section ??.



(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)

100. Information gain is relatively constant after 130 topics. Based on this measure, the 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:

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

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 section 5. We assign our schema using our machine learning methodology. After deriving our 120 topics, we assigned one class per topic.²¹ 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.²² 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

²¹Where 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.

²²We 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 in section 2.1. 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 1617-1850.²³ 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.

²³While 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*** (0.065)	0.189*** (0.045)	- -	0.153*** (0.046)	0.099*** (0.023)	0.119*** (0.015)
Clothing	0.061 (0.046)	0.051 (0.040)	0.011 (0.041)	0.102** (0.040)	0.016 (0.035)	0.054* (0.031)
Engines	0.214*** (0.037)	0.183*** (0.040)	- -	0.166*** (0.043)	0.081*** (0.028)	0.110*** (0.019)
Food	- -	0.118*** (0.042)	0.166*** (0.050)	-0.067 (0.103)	0.040 (0.052)	-0.015 (0.032)
Instruments	- -	0.044 (0.055)	-0.026 (0.059)	-0.017 (0.058)	-0.015 (0.041)	-0.036 (0.026)
Medicines	0.271*** (0.040)	0.242*** (0.036)	0.237*** (0.043)	0.219*** (0.038)	0.133*** (0.030)	0.179*** (0.027)
Metal	0.178*** (0.058)	0.164*** (0.052)	- -	0.137*** (0.051)	0.069 (0.043)	0.086*** (0.030)
Military	-0.064 (0.051)	-0.028 (0.064)	-0.048 (0.064)	-0.114* (0.067)	-0.126** (0.053)	-0.105*** (0.034)
Mining	0.261*** (0.081)	0.210*** (0.066)	0.206*** (0.060)	0.148*** (0.046)	0.110*** (0.038)	0.104*** (0.022)
Paper	0.341*** (0.052)	0.130*** (0.043)	0.083** (0.040)	0.064 (0.046)	-0.028 (0.026)	0.016 (0.012)
Textiles	-0.042 (0.056)	-0.033 (0.075)	-0.006 (0.067)	-0.050 (0.055)	-0.063 (0.054)	-0.046* (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 (0.187)	0.245 (0.150)	- -	0.078 (0.099)	0.351*** (0.092)	-0.002 (0.066)
Clothing	-0.429*** (0.136)	0.200 (0.273)	-0.302* (0.161)	-0.369* (0.201)	0.109 (0.285)	-0.399*** (0.129)
Engines	0.085 (0.175)	0.244 (0.191)	- -	0.188* (0.096)	0.298** (0.126)	0.072 (0.073)
Food	- -	0.233 (0.166)	-0.138 (0.128)	-0.029 (0.247)	-0.036 (0.206)	-0.265 (0.182)
Instruments	- -	0.216* (0.116)	0.175 (0.153)	0.105 (0.071)	0.251** (0.106)	-0.022 (0.047)
Medicines	-0.476*** (0.137)	-0.282** (0.117)	-0.495*** (0.127)	-0.381 (0.324)	0.004 (0.338)	-0.336 (0.298)
Metallurgy	0.093 (0.187)	0.355** (0.166)	- -	0.205* (0.118)	0.374*** (0.107)	0.149 (0.118)
Military	0.151 (0.152)	0.455** (0.188)	0.173 (0.148)	0.378*** (0.113)	0.290* (0.148)	0.132 (0.119)
Mining	0.606 (0.415)	0.367 (0.235)	0.075 (0.195)	0.238 (0.161)	-0.011 (0.119)	-0.176*** (0.057)
Paper	-0.078 (0.302)	0.310* (0.164)	0.244 (0.162)	0.206* (0.115)	0.221* (0.114)	0.032 (0.091)
Textiles	0.404* (0.234)	0.561** (0.218)	0.347* (0.205)	0.448** (0.189)	0.475*** (0.128)	0.291** (0.146)
Constant	-0.480 (0.430)	-0.805* (0.451)	-0.605 (0.389)	-0.650* (0.363)	-0.777** (0.374)	-0.590 (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 (0.082)	-0.019 (0.044)	- (0.043)	0.033 (0.034)	0.022 (0.022)	-0.010 (0.022)
Clothing	0.055 (0.056)	-0.003 (0.041)	-0.044 (0.043)	-0.002 (0.049)	-0.043 (0.031)	-0.077*** (0.027)
Engines	0.001 (0.039)	-0.005 (0.041)	- (0.043)	0.008 (0.032)	0.017 (0.021)	-0.022* (0.013)
Food	- (0.039)	-0.029 (0.039)	-0.037 (0.039)	-0.061 (0.058)	0.074 (0.079)	-0.026 (0.053)
Instruments	- (0.040)	-0.041 (0.040)	-0.033 (0.041)	-0.002 (0.034)	-0.014 (0.020)	-0.054*** (0.013)
Medicines	-0.077* (0.041)	-0.110*** (0.040)	-0.089** (0.039)	-0.115*** (0.034)	0.018 (0.067)	-0.063 (0.048)
Metal	0.049 (0.053)	0.011 (0.041)	- (0.048)	0.070 (0.048)	0.073* (0.039)	0.037 (0.033)
Military	-0.087* (0.052)	-0.095** (0.043)	-0.080 (0.049)	-0.060* (0.035)	-0.076* (0.044)	-0.105*** (0.023)
Mining	-0.083 (0.069)	0.048 (0.080)	-0.008 (0.054)	0.030 (0.047)	0.015 (0.036)	-0.014 (0.022)
Paper	-0.018 (0.044)	-0.028 (0.050)	-0.005 (0.048)	0.012 (0.035)	-0.031 (0.034)	-0.052** (0.022)
Textiles	0.125*** (0.041)	0.112** (0.046)	0.070 (0.045)	0.099** (0.044)	0.095*** (0.029)	0.051** (0.025)
Constant	0.018 (0.070)	0.007 (0.056)	0.009 (0.068)	-0.015 (0.067)	-0.022 (0.056)	0.028 (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 (0.046)	-0.039 (0.030)	- -	-0.017 (0.025)	0.001 (0.027)	-0.112*** (0.015)
Clothing	0.448*** (0.043)	0.228*** (0.039)	0.253*** (0.049)	0.237*** (0.062)	0.213*** (0.048)	0.082* (0.048)
Engines	-0.018 (0.030)	-0.002 (0.034)	- -	0.044 (0.094)	-0.018 (0.097)	-0.080 (0.051)
Food	- -	-0.075*** (0.029)	-0.018 (0.042)	-0.071 (0.059)	-0.055 (0.034)	-0.124*** (0.037)
Instruments	- -	-0.035 (0.051)	0.012 (0.061)	0.041 (0.048)	0.007 (0.035)	-0.030 (0.036)
Medicines	0.097 (0.070)	0.062 (0.069)	0.146** (0.065)	0.139*** (0.045)	0.117*** (0.039)	0.057* (0.034)
Metal	0.133** (0.063)	0.146*** (0.053)	- -	0.139* (0.072)	0.034 (0.069)	0.013 (0.055)
Military	0.111* (0.057)	0.070 (0.061)	0.101* (0.059)	0.082** (0.039)	-0.014 (0.039)	-0.013 (0.030)
Mining	0.058 (0.084)	0.110* (0.056)	0.141** (0.060)	0.197*** (0.038)	0.145*** (0.031)	0.085*** (0.027)
Paper	0.348*** (0.065)	0.177*** (0.040)	0.173*** (0.039)	0.035 (0.036)	0.029 (0.026)	-0.028 (0.028)
Textiles	0.393*** (0.040)	0.318*** (0.035)	0.306*** (0.038)	0.271*** (0.040)	0.245*** (0.049)	0.174*** (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 section 5 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 used in section 5.²⁴ 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:

$$NT_{ci} * Metric_i = \alpha_i + \beta S_{ci} * Metric_i + \mu_i \quad (4)$$

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 5. 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.

²⁴We 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 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

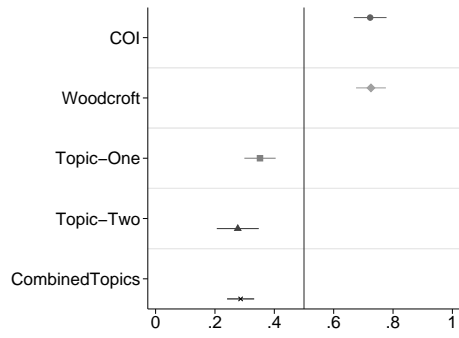
are strongly correlated with the NT schema. By contrast, the Topic-One, Topic-Two, and CombinedTopics schemas are much less correlated with NT. To understand whether regression plots can be useful for predicting possible divergence, we contrast the plots for another patent class where the COI or Woodcroft 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.

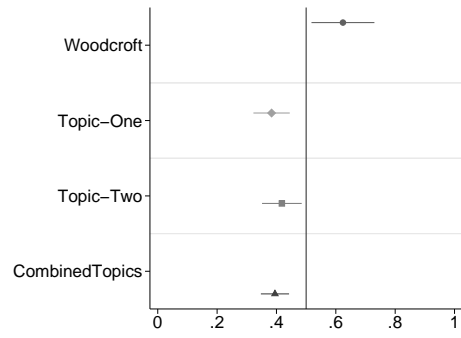
Appendix E.2 – Capital Saving Patents

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.

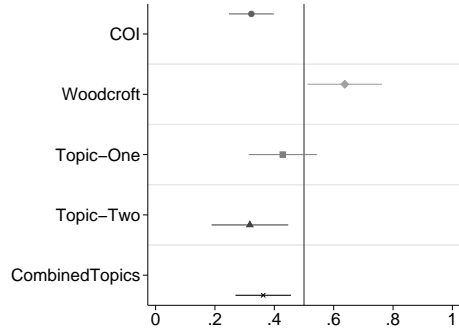
Figure E2 reports the regression plots for the capital-saving variable. Compared to the patent quality regression plots, there is a stark difference when observing capital-saving patents. While the general trend remains similar to the patent quality metric, the confidence intervals associated with capital-saving coefficients for all of the



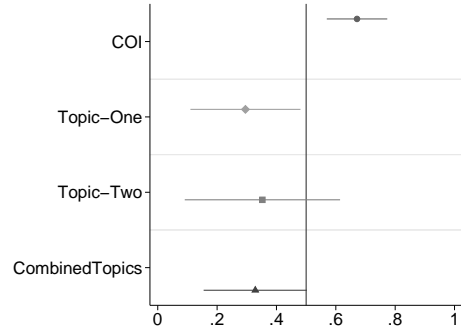
(a) Agricultural patents



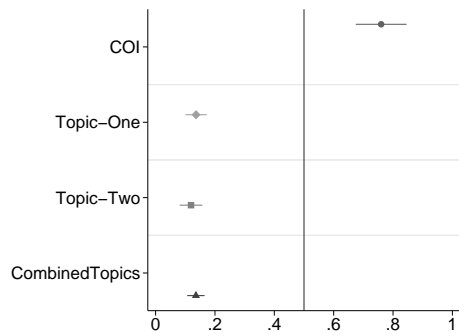
(b) Chemical patents



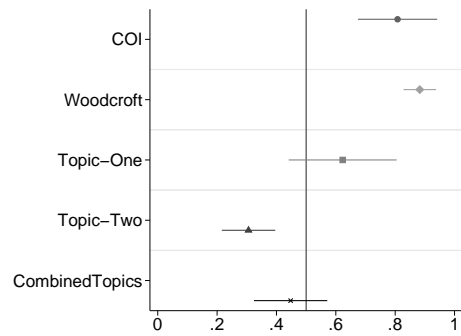
(c) Clothing patents



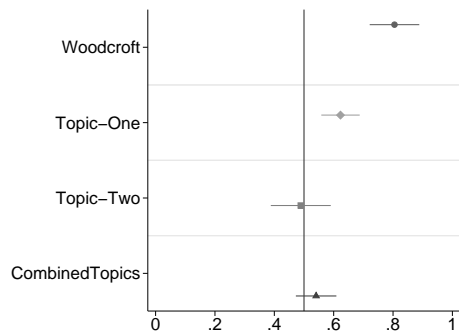
(d) Food patents



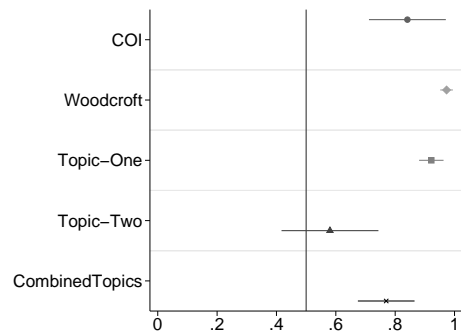
(e) Instrument patents



(f) Medicine patents



(g) Metal patents



(h) Military patents

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 patent class are also dummies, while patent quality is a continuous numerical variable.

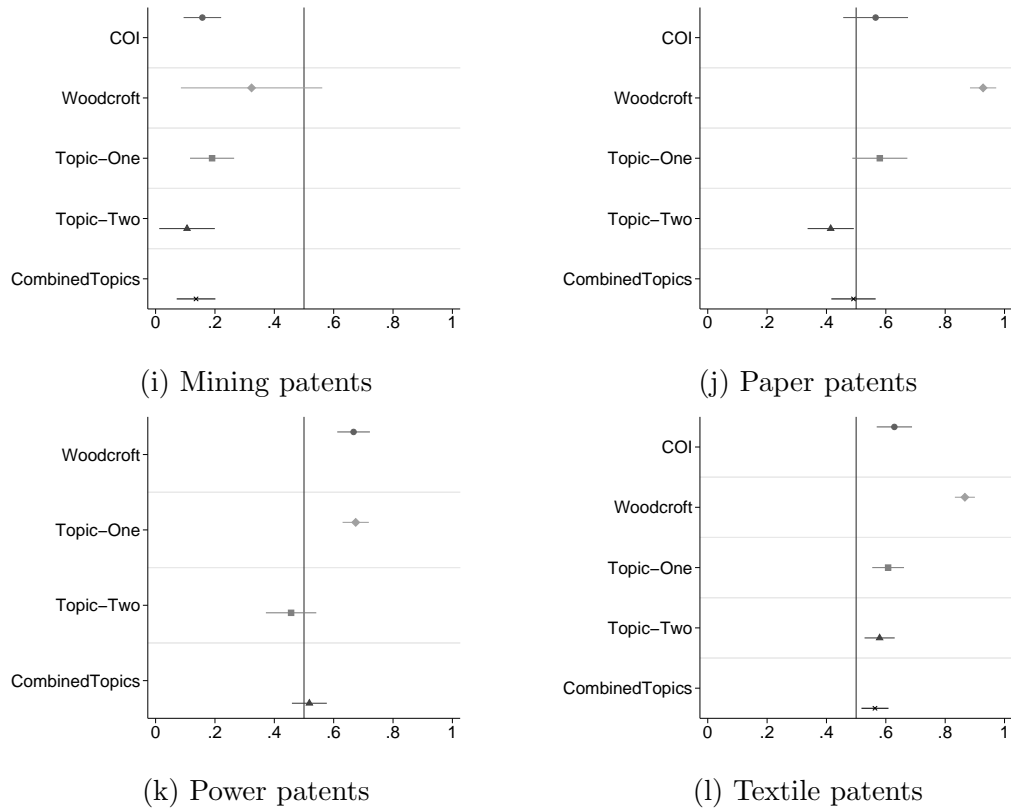
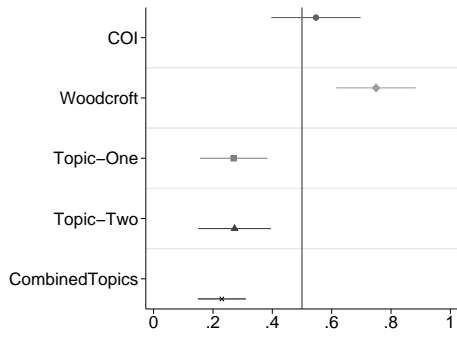


Figure E1: Regression plots for the citations of patented inventions

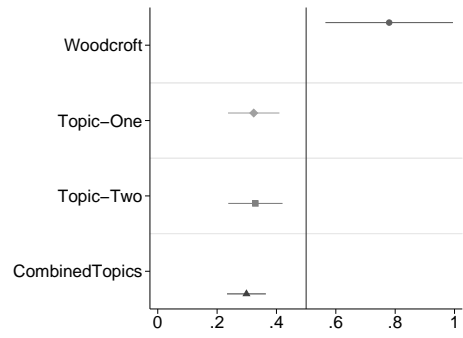
Notes: The plots depict the results from OLS regressions of the Nuvolari-Tartari (NT) schema against the Cradle of Invention (COI), Woodcroft, Topic-One, Topic-Two, and CombinedTopics schemas. The variable of interest is the Woodcroft Reference Index. Each plot represents one of the 12 common patent classes analysed in section 5. 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.

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.

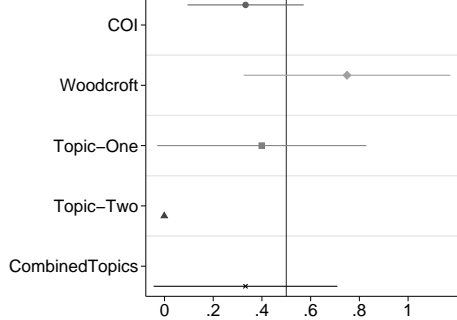
Comparing these results with the regressions in Table 9 may highlight whether the regression plots could predict the likely degree of classification divergence. Similar to patent quality, the COI and Woodcroft schemas are correlated with NT, so we may expect



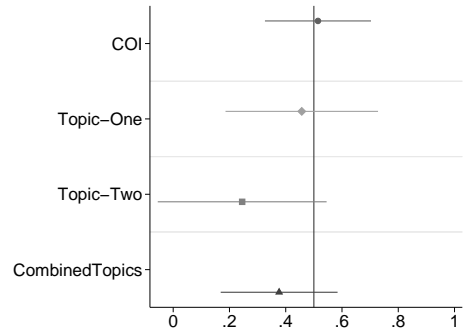
(a) Agricultural patents



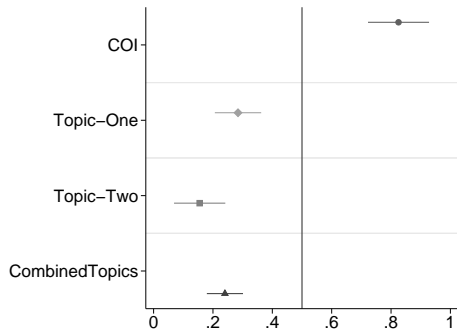
(b) Chemical patents



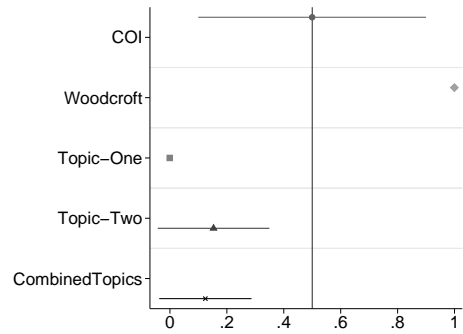
(c) Clothing patents



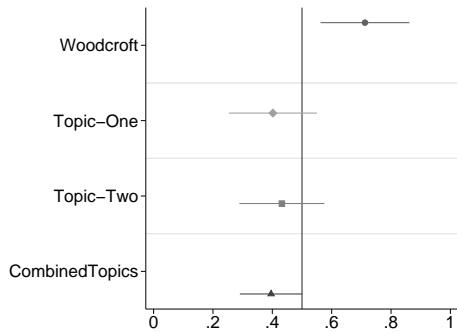
(d) Food patents



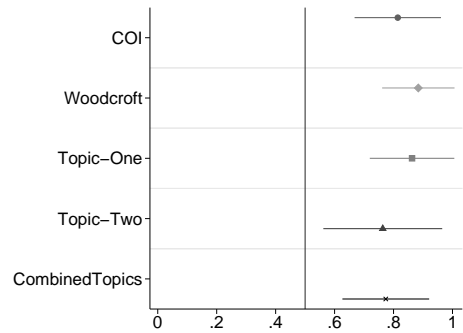
(e) Instrument patents



(f) Medicine patents



(g) Metal patents



(h) Military patents

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 patents, therefore we may expect to see similar results for Chemical patents in Table 9

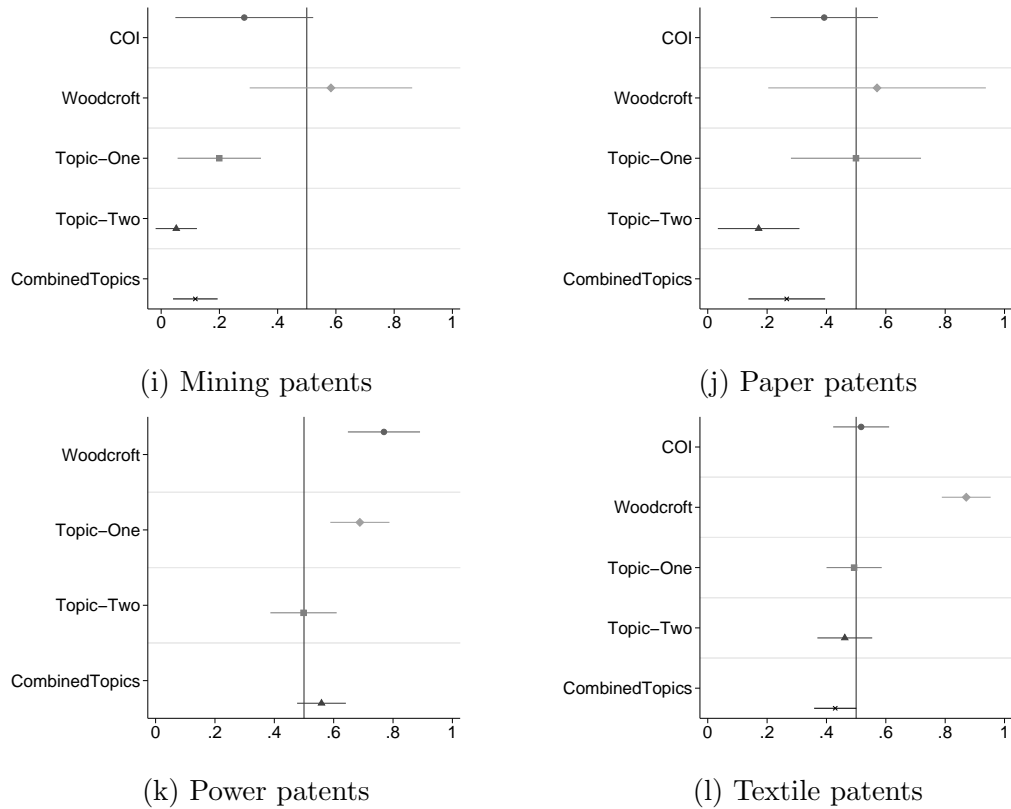


Figure E2: Regression plots for capital saving patents

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. 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.

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 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.