RIETI Discussion Paper Series 20-E-010

Science and Technology Co-evolution in AI: Empirical Understanding through a Linked Dataset of Scientific Articles and Patents

MOTOHASHI, Kazuyuki
RIETI

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 20-E-010
February 2020

Science and Technology Co-evolution in AI: Empirical understanding through a linked dataset of scientific articles and patents*

By Kazuyuki Motohashi

(University of Tokyo, NISTEP and RIETI, Japan)

Abstract

The linked dataset of AI research articles and patents reveals that a substantial public sector contribution is found for AI development. In addition, the role of researchers who are involved both in publication and patent activities, particularly in the private sector, increased over time. That is, open science that is publicly available through research articles and propriety technology that is protected by patents are intertwined in AI development. In addition, the impact of data science, measured by AI research articles on innovation, is analyzed by patent citation analysis. It is found that patents invented by AI paper authors are more likely to have more forward citations by other applicants (non-self-citation), in wider technology fields (greater generality index). This implies that the nature of general purpose technology (GPT) for data science is elevated by the fact that patent inventors are also involved with scientific activities and published as research authors.

Keywords: Artificial Intelligence (AI), bibliometric analysis, patent data

JEL classification: O31, O3

The RIETI Discussion Papers Series aims at widely disseminating research results in the form of professional papers, with the goal of stimulating lively discussion. The views expressed in the papers are solely those of the author(s), and neither represent those of the organization(s) to which the author(s) belong(s) nor the Research Institute of Economy, Trade and Industry.

* This study is conducted as a part of the Project "Digitalization and Innovation Ecosystem: A Holistic approach" undertaken at the Research Institute of Economy, Trade and Industry (RIETI).

1. Introduction

AI (machine learning) is perceived as a key technology causing fundamental changes of innovation landscape by making future prediction cheaper and more certain (Agrawal et al., 2018). Therefore, private incentives to capture such potential values are substantial, and huge attention to this technology is found both in private firms and academic sector. Another characteristics of AI is that its potential applications can be found in variety of industry. In other words, AI is a killer application of IT as general purpose technology (Helpman, 1998). Furthermore, AI can be served as a new method of invention (IMI: Invention as a Method of Inventing), as is seen in AI use for new drug discovery (Cockburn et al., 2018).

It is also found that the co-occurrence of publication and patenting at individual engineer level is very popular in this field, suggesting that the co-evolution of science and innovation is happening (Motohashi, 2018). For example, the deep learning (deep neural network) is used in various industrial application, but an initial implementation of new methodology was made by academia. Subsequently, a series of development of deep learning algorithms, suited for various kinds of datasets (such as CNN for image data and RNN for text data) are developed by computer science scholars. A substantial contribution to AI development by private sector is also found. A typical case example is Google Brain's publication of "alpha go" (Silver et al., 2017). A team at Deep Mind, currently under the Google AI department (Google Brain), not only developed the software to beat the world go (Chinese chess) champion, but also made it in a public as a research paper. A shorter distance between science and innovation is found in many fields, where a shift from linear model (science -> innovation) to co-evolution of them are found in open and digital era (OECD, 2019).

This paper sheds new light on the nature of AI (machine leaning) focusing on the interactions between scientific publication and patenting. The research questions include "why a firm publish freely their new findings in the field of AI, as well as patenting some of them as propriety technology?" and "can such behavior be explained by using business ecosystem concept, in a sense that a key stone player needs to balance providing its managerial resource to whole ecosystem players and appropriating the economic rents?". We use the linked dataset of SCOPUS research article database and USPTO patent information at author/inventor level. There are around 8 million papers from SCOPUS and 3 million patents from USPTO data. These two data are linked by author/inventor names as well as his/her affiliates, and about 5% of all authors from SCOPUS and about 13.3% of inventors from USPTO data can be linked (Motohashi, 2018). We have constructed the AI patent datasets, centered on the IPC subclass "G06N" (WIPO, 2019; JPO, 2019). Then, the empirical analysis is conducted the impact of being a paper author as an

inventor of such patent on subsequent innovations, measured by patent citation information.

The next section described the concept of AI (data science) driven innovation. Here the interrelationship between AI and the complementary elements to innovation, i.e., big data and IoT is discussed. Then, an analytical part by using the linked dataset of research articles and patents is provided. This section is followed by the statistical analysis using patent citation information. Finally, this paper concludes with some managerial and policy implications.

2. AI as a driver of data driven innovation

Faster and cheaper computer power and internet environment open new opportunities to use computer in more intelligent way, such as recognition, inference and future prediction. An application of AI includes image or text data recognition technologies, human interfaces (visualization of data and interactive agents), knowledge discovering technologies related to the diagnosis, monitoring, and datamining of various types of equipment devices. Combining these components (enablers of downstream innovation) lead to various industrial applications, called smart XXX (XXX includes home appliances, factory, energy, maintenance, and medical services).

One of essential components of AI is data (or called big data), used for machine learning. Regardless of whether supervised or unsupervised learning, the 3Vs (Volume, Variety and Velocity) of big data are important. In supervised learning, a large volume of the text and image data accumulated on the internet can be used as training data. For example, Google provides translation services by having their translation system read a large volume of documents written in two or more languages (training data) to construct translation models. Conventional machine translation systems utilize rule-based models, which is based on grammatical structure of sentences and word dictionary. On the other hand, in the models that use machine learning, computers produce rules for translation from a large volume of documents (corresponding documents between English and Japanese for Japanese-English translation, for example), which are provided as inputs. In other words, computers automatically perform language parsing work, which is the basis for translation rules in place of the one developed by linguists. In this case, since human thinking is replaced by a computer, we can consider this to also be one example of AI.

Another important component of AI is IoT, consisted by sensor and network technologies. A tremendous amount of data is generated by various sensors around us. The realization of IoT requires various elements—identification, or the labelling of each item using an IP address; sensing, or the measurement and “datafication” of the item; communication, which primarily involves data

communication; computation, or an analysis of the item's data—and its implementation as a specific service, such as the maintenance and operation of industrial machinery or buildings' energy management systems (Al-Faqaha et al., 2015). It is believed that one trillion things, or 100 times the human population, could be connected by the year 2020. As networking expands from people to "things," the data volume will also dramatically increase. As it is unrealistic to exchange all information through the Internet, "edge" computing has also attracted attention, as this forms local networks and performs distributed processing. This can enable more expanded applications through the aggregation of a certain level of information from those local networks and the connection of "things" in wider areas through the Internet. Consequently, information of all kinds will become connected worldwide through the Internet.

(Figure 1)

AI plays an important role to convert such data to various industrial applications or data driven innovation (smart XXX), with leaning (knowledge accumulation), inference and prediction function. Deep learning, a method of machine learning that uses a multi-layered neural network, tremendously improves prediction accuracy. A neural network is a classical mathematical method with decades of history. There have conventional ideas of deep learning to make a multilayered network layer; however, there was a problem in that it was difficult to estimate parameters, which increase through the creation of a multilayered network. In addition, the abilities and performance of a computer are insufficient. In recent years, deep learning has been re-examined, and AI studies are now hot spots since computer performance has been improved, and a large volume of information has been compiled on the internet, which enables big data to be utilized when estimating models. In recent years, estimation methods have been developed for respective types of data (e.g., image data or text data) and characters, and implemented in various fields, including industrial applications such as industrial robots and autonomous operation technologies, investment decision-making for financial institutions, financial advisory work, and household appliances such as cleaning robots and AI speakers.

3. Measuring AI driven innovation by patent and research article information

We have linked patent and research article data at author/inventor level, more specifically, finding identical author/inventor in the database of research articles and patents in order to measure AI driven innovation presented in the previous section. We focus on some important features of AI, such as machine learning, driving large numbers of downstream innovations (smart XXX). Specifically, we use the framework developed by WIPO to characterize the nature of AI innovation by patent and research article information (WIPO, 2019).

The WIPO's analytical framework is constructed by three layers of AI, methodology (such as machine learning), function (such as image recognition) and applications (such as autonomous driving). As regards to AI related patents, the key IPC subclass is G06N, representing "computer systems based on specific computational models, such as neural network, inference machine and fuzzy logic". This classification is used for "methodology" layer of WIPO framework, and also used by IP offices for patent based AI technology review (JPO, 2014; JPO, 2019). WIPO have retrieved patents related to the other two layers of its framework ("function" and "application") by using other IPC classes and keywords. However, it is shown that WIPO's three layer framework can be reconstructed by using "G06N" as a basis together with the other IPC information given to the same invention (Motohashi, 2019). Therefore, we use this IPC subclass as a definition of AT patents in subsequent analysis.

In terms of research article information, we use SCOPUS database by Elsevier, where ASJC (All Science Journal Classification) is given in each journal. Here, ASJC 1702 is labeled as "Artificial Intelligence" under a broad category of "computer science", which is used in this paper as AI paper identification. It should be noted that ASJC classification is made at journal level, instead of individual paper level, so that AI paper extraction by ASJC code may not capture the emerging trend of AI for non-AI related journals. Therefore, the robustness check of AI paper trend is conducted by using keyword matching, used in Coburn et al. (2017), and confirmed that the trend of these two sources do not give very different results.

Figure 2 shows the trend of the shares of AI papers (with ASJC 1702 and US based affiliated organization) and patents (USPTO patents with G06N subclasses) to the all of discipline. Both of them have upward trend until 2010, but stable afterwards. It should be noted USPTO discloses the information of only granted patents. Therefore, only data applied until 2011 can be used even the datasets are obtained from USPTO data download site, called patentview.org in 2016. In Figure 2, the trend of AI patent applications, whose truncation bias is relatively small, are also presented. The surge of AI patent applications after 2010 is found. It should be also noted that the shares of AI papers and patents are very small, like less than 1% of total papers and patents, since only core technology of AI is included in both definitions.

(Figure 2)

We, then, linked the USPTO patent and SCOPUS research article at individual researcher level, to investigate science and technology coevolution of AI. In both datasets, we select the researchers working for the organizations located in the United States. There are around 8 million papers from SCOPUS and 3 million patents from USPTO data. These two data are linked by author/inventor names as well as his/her affiliates, and about 5% of all authors from

SCOPUS and about 13.3% of inventors from USPTO data can be linked (Motohashi, 2018). This approach is taken from the one of measuring science and innovation co-evolution in Japan (Ikeuchi et al., 2016). A similar study is found in constructing the matched data of paper and patent pair by their contents (Lissoni et al., 2013). But, we take a broader context of science and innovation co-occurrence at engineer level, even the contents of the two sources are difference. Traditionally, the degree of scientific basis, or 'science intensity' of industry has been measured using non-patent literature (research article) citations made by patents (Narin and Noma 1985, Schmoch, 1997). Non-patent literature citations show the degree of disembodied scientific knowledge that flows into patents, while the patent-publication pair can capture the state of co-occurrence of scientific and invention activities within the same researchers, i.e., interplay of science and technology embodied in human capital.

Next, we look at the contribution of AI author/inventor contribution to aggregated trend of AI papers and patents, and it is found that the shares of AI papers and AI patents of such cross over researchers are greater than those of pure authors and pure inventors, respectively (Figure 3). The difference of these two groups is particularly large in AI patent shares. That is, more and more appropriation of AI technology by patents are observed in AI scientist, who also contributed to research article publication activities.

(Figure 3)

4. Statistical analysis: S&T co-evolution at researcher level

Since our focus of statistical analysis is to evaluate S&T co-evolution at individual researcher level, we start with AI authors (a researcher who has at least one AI paper as the definition above) and extract all patents invented those researchers. Those patents are compared with the patents invented by AI inventors (an inventor who has at least one AI patent as the definition above) to see the impact of science linkage on their subsequent inventions. Figure 4 shows the difference of technology classification (WIPO's technology classification of 35 categories) of them (patents by AI authors and AI inventors). In Figure 4, we take out the category of "computer technology" where G06N (AI patents) is included. [1] It is found that the patents invented by AI authors have wider applications across technology field, as compared to those by AI inventors. The difference between two figures are found particularly in "measurement", "medical technology", "transport" and "organic fine chemistry". These findings reflect that the AI inventors with scientific activities measured by publication contributes significantly to the

[1] The share of computer technology is 53.2% for AI inventors and 41.6% for AI authors.

AI's nature of GPT (general purpose nature) and/or IMI (invention of method of invention).

(Figure 4)

Another dimension of S&T co-evolution is related to inventor's affiliated organization type. Since our analysis is based on disambiguated inventor information at USPTO official website (http://www.patentview.org/), it is possible for us to identify "cross-over" inventor who moves across scientific sector (university or public research institution) and industry (private firm).[2] It is found that the share of patents of crossover inventor is particularly high in AI patents (Figure 5). More than 25% of AI patents have at least one crossover inventors, as compare to its average value of less than 10%.

(Figure 5)

In order to see the impact of science and technology co-evolution in subsequent inventions, a regression analysis is conducted to compare the patent invented by AI author to that invented by AI inventor (but not involving with research article as an author). We use the patents invented after 2000 for the regression analysis, and the number of samples of AI authors patents group is 51,946, while that of AI inventors are 75,252. The dependent variables are the number of forward citations, the number of non-self forward citations and generality index (Trajtenberg et al., 1998). The key explanatory variable is a dummy of invented by AI author or not, as well as followings,

- A dummy for a patent invented by the author of AI papers
- A dummy of NPL citation
- A dummy for public sector as a patent assignee (1 for academia and 0 for private firm)
- A dummy with patent invented by at least one crossover inventor between academia and private firm
- Interaction terms of AI paper dummy with NPL, public and crossover
- A dummy with patent applied after AI paper published (used for AI author patent samples only)

[2] Machine learning technique is used to identify patents with an identical inventor, based on USPTO inventor records with synonym problem (in which the same person's name appears in several distinct forms due to name changing etc.) and homonym problem (in which many distinct people share the same name). A survey of inventor disambiguation works comparing machine learning methodologies is found in Yin el. al (2019).

We also control for application year and IPC subclasses of each patent, and the regression analysis are conducted for all samples to control for cross time and technology difference of citation indicators. OLS regression results are presented in Table 1-1, 1-2 and 1.3 for three types of dependent variables.

(Table 1-1), (Table 1-2), (Table 1-3)

In both total citation and non-self citation, a dummy of AI publication (invented by AI author) has positive and statistically significant coefficients in all models. It should be noted that this result is robust even after controlling for NPL dummy (citing to non-patent literature, reflecting science linkage of patent content). In terms of the role crossover inventor in subsequent patent citations, the mixed results are found. That is, negative and statistically significant coefficient is found in the models (Table 1-1 and Table 1-2). However, the coefficient to the interaction term with AI publication has positive and greater absolute value, so that the total impact of crossover inventor should be positive (2.199-0.599>0 in Table 1-1, for example). Or, this should be interpreted by crossover inventor dummy being negatively correlated for the controlling samples, while positively correlated for the samples with AI publication authors.

In addition, we have conducted the regression analysis, only for AI author samples, by including a dummy with patent applied after AI paper published (mode (5) in all three tables). It is found that a dummy for the invention after AI publication has positive and statistically significant coefficients. Therefore, the impact of AI publication for subsequent inventions are confirmed even after controlling for the inventor level characteristics (such as research quality and social capital by researcher networks).

In terms of generality index, reflecting technological diversity of citing patents, a dummy of AI publication leads to greater values in general (Table 1-3 and Table 2-3). A positive association of crossover inventor, as well as its complementary relationship with AI publication is also found. However, a different result from those of forward citations is found in the coefficient to a dummy for invention after publication. That is, the generality is lower for the patents invented after AI publication made. Taken together with the results of forward citation, the patents invented after AI publication is more likely to be cited by other patents, but by those with narrower technology focus.

5. Robustness check

The empirical results presented in the previous literature depends on the comparability of controlling samples to the treated group (patents by AI authors). Therefore, we have conducted a

robustness check by using different kind of controlling group and redo the regressions. We use matching samples created by the same IPC subgroup and application year with the samples in treated group, instead of AI inventor patents in the previous section. The sample size of this control group is 40,201 (type 2), while it is the same as above for the treated group (51,946). The results are presented in Table 2-1, 2-2 and 2-3.

(Table 2-1), (Table 2-2), (Table 2-3)

The results do not change from the previous ones. A one difference is that coefficient to cross over inventor becomes to be positive and significant, as we as positive coefficient to the interaction term with AI author paper dummy. However, this does not change a key story here, that is, positive impact of AI publication to subsequent inventions and its complementarity with a talent cross over public and private institutions.

6. Discussion and implications

In this paper, coevolution of science and technology in AI field is investigated by the linked dataset of AI research articles and patents at individual researcher level. It is found that the interaction between patent and research article occurs more frequently in AI field, and such trend increases over time, particularly in the share of AI author patents. In addition, the share of inventor moving across academic and industry sector (cross over inventor) is higher in AI patents. This is, open science, publicly available by research articles and propriety technology, protected by patents are intertwined in AI development.

Furthermore, our regression analysis reveal that the patent invented by AI paper author have more impacts on subsequent inventions, both in self citation and non-self citation, and the generality index of forward citation is greater, as compared to the patents with similar contents. The existence of cross over inventors positively moderates such relationship, so that the impact of publication on subsequent inventions are reinforced by the inventor who has working experience both in academic and industry sectors.

As regards to the original research question of this paper, “why does a private open up its technological findings as research article publication?”, our empirical findings support the view of eco-system building, in a sense that a keystone player (or a platformer) in business ecosystem is supposed to provide its managerial resources to niche players (or platform users) in order to maximize the value of whole ecosystem (Iansiti and Levien, 2004). Opening up of technology as a form of publication as well as making related technologies be proprietary one by patent, leads to higher subsequent inventions such technologies. Subsequently, a firm with publication

activities are able to create greater numbers of followers of the firm's technology stream,

Another potential answer to the question of why a firm publishes is based on the requirement of accessing talents in academic sector. Due to explosion of demand for AI and data analytics works, labor market for data scientists becomes extremely tight. Therefore, it is important for a firm to offer attractive working environment for them. One of incentives for them to work for a private firm is higher salary and access to its internal propriety big data. However, financial incentive may not be enough to attract a top notch data scientist, so that some of firms give some opportunity for its employee to work on her own project, and academic activities such as participating in academic conferences and publications. In our works, it is found that cross over talents between academia and industry plays complementary role in positive relationship between publication and subsequent inventions. This finding is consistent with the view of human resource reason of private firm's publication, in sense that a firm supposed to offer some room for academic activities to those academic researchers who are capable to conduct high impact research and engineering activities at the firm.

Managerial implications from our empirical study is directly related to foregoing discussion. First, it is important to understand the nature of AI driven innovation, active science and technology co-evolution, in order to tap on huge economic opportunities by using such new technologies. Therefore, a good balance between open and close strategy as regards to the outputs of technological activities are important. Second, understanding the incentives of academic researchers are important. In AI field, university and industry collaboration activities should involve substantial interactions of human resources across industry and academia. Therefore, a firm should not enforce too tight regulation over joint activities with academic, but it may be more effective if a firm allows researchers involved in such joint activities to publish their research findings.

References

Al-Fuqaha, A., Guizani, M. Mohammadi, M., Aledhari, M. and M. Ayyash (2015), Internet of Things: A Survey on Enabling Technologies, Protocols and Applications, IEEE Communications Surveys and Tutorials

Agrawal. A., J. Gans and A. Goldfarb (2018), Prediction Machines: The Simple Economics of Artificial Intelligence, Harvard Business School Press, April 2018

Arora A., S. Belenzon and A. Patacconi (2015), Killing the Golden Goose? The Decline of Science in Corporate R&D, NBER Working Paper #20902

Cockburn, I., Henderson, R. and S. Stern (2018), The impact of artificial intelligence on innovation, NBER working paper #24449, March 2018

Helpman, E (1998), *General purpose technologies and economic growth*, MIT Press, Cambridge MA

Iansiti,, M. and R. Levien (2004), The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability, Harvard Business School Press, Boston, MA

Ikeuchi, K, K Motohashi, R Tamura, and N Tsukada (2017), "Measuring Science Intensity of Industry using Linked Dataset of Science, Technology and Industry", RIETI Discussion Papers Series 17-E-056.

JPO (2019), Survey of AI related patent applications, Japan Patent Office, July 2019 (in Japanese)

Lissoni, F, F Montabio, and L Zirulia (2013), "Inventorship and authorship as attribution rights: an enquiry into the economics of scientific credit", Journal of Economic Behavior and Organization, 95, 49-69

Motohashi, K. (2019), Measuring AI innovation and science linkage by patent data, Keizai Tokei Kenkyu, 47(2), Economy, Trade and Industry Statistics Association, Tokyo Japan (in Japanese)

Motohashi, K. (2018), Co-occurrence of science and innovation in AI : Empirical analysis of paper-patent linked dataset in the United States, NISTEP working paper No. 160 (in Japanese)

Narin, F, and E Noma (1985), "Is technology becoming science?", Scientometrics, 7, 368-381.

OECD (2019), *Digital Innovation: Seizing Policy Opportunities*, OECD, April 2019, Paris France

Silver D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel and D. Hassabis (2017), Mastering the game of Go without human knowledge, Nature Vol 55, 19 October 2017

WIPO (2019), Technology Trend 2019: Artificial Intelligence, Data collection method and

clustering scheme, WIPO back ground paper, Geneva, Switzerland

Yin, D., K. Motohashi and J. Dang (2019), Large-scale name disambiguation of Chinese patent inventors (1985-2016), *Scientometrics*, published on-line, 1-26

Table 1-1: Regression results (Forward Citation, vs AI Patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	With autor
With paper author dummy	0.848 [0.084]**	0.816 [0.140]**	0.861 [0.084]**	0.349 [0.148]*	
NPL dummy	1.826 [0.087]**	1.821 [0.108]**	1.831 [0.087]**	1.762 [0.108]**	1.974 [0.153]**
Public dummy	0.949 [0.314]**	-0.458 [0.670]	0.872 [0.320]**	-0.073 [0.673]	0.355 [0.402]
With Author *NPL		0.016 [0.172]		0.085 [0.172]	
With Author*Public		1.794 [0.755]*		0.295 [0.768]	
With Crossover Inventor			0.129 [0.100]	-0.599 [0.121]**	1.866 [0.246]**
With Author*Crossover				2.199 [0.205]**	
After publication					0.421 [0.155]**
After pub*Crossover					-0.404 [0.344]
Constant	-0.769 [0.776]	-0.753 [0.777]	-0.812 [0.776]	-0.522 [0.777]	-1.406 [1.317]
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.13	0.13	0.16
# of observations	127,198	127,198	127,198	127,198	51,946

**: Statistically significant at 1% level, * Statistically significant at 5% level

Table 1-2: Regression results (Non-self Forward Citation, vs AI patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	With autor
With paper author dummy	0.998 [0.076]**	0.835 [0.127]**	1.023 [0.077]**	0.474 [0.135]**	
NPL dummy	1.396 [0.079]**	1.314 [0.098]**	1.405 [0.079]**	1.277 [0.098]**	1.655 [0.144]**
Public dummy	1.389 [0.285]**	0.234 [0.608]	1.240 [0.290]**	0.482 [0.611]	0.804 [0.378]*
With Author *NPL		0.224 [0.156]		0.265 [0.156]	
With Author*Public		1.465 [0.686]*		0.218 [0.697]	
With Crossover Inventor			0.249 [0.090]**	-0.358 [0.109]**	1.698 [0.232]**
With Author*Crossover				1.810 [0.186]**	
After publication					0.357 [0.146]*
After pub*Crossover					-0.417 [0.323]
Constant	-0.715 [0.704]	-0.661 [0.705]	-0.798 [0.705]	-0.519 [0.706]	-1.334 [1.240]
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.13	0.13	0.15
# of observations	127,198	127,198	127,198	127,198	51,946

**: Statistically significant at 1% level, * Statistically significant at 5% level

Table 1-3: Regression results (Generality Index, vs AI Patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	With autor
With paper author dummy	0.024 [0.002]**	0.032 [0.003]**	0.023 [0.002]**	0.016 [0.003]**	
NPL dummy	0.027 [0.002]**	0.032 [0.002]**	0.027 [0.002]**	0.030 [0.002]**	0.024 [0.003]**
Public dummy	0.052 [0.007]**	-0.007 [0.016]	0.057 [0.007]**	0.014 [0.016]	0.036 [0.008]**
With Author *NPL		-0.013 [0.004]**		-0.010 [0.004]**	
With Author*Public		0.072 [0.018]**		0.025 [0.018]	
With Crossover Inventor			-0.008 [0.002]**	-0.033 [0.003]**	0.046 [0.005]**
With Author*Crossover				0.068 [0.004]**	
After publication					0.004 [0.003]
After pub*Crossover					-0.030 [0.007]**
Constant	0.099 [0.022]**	0.096 [0.022]**	0.101 [0.022]**	0.109 [0.022]**	0.109 [0.034]**
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.13	0.13	0.17
# of observations	84,313	84,313	84,313	84,313	36,872

**: Statistically significant at 1% level, * Statistically significant at 5% level

Table 2-1: Regression results (Forward Citation, vs Same IPC Patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	ALL
With paper author dummy	1.454 [0.099]**	1.305 [0.161]**	1.309 [0.100]**	1.066 [0.165]**	1.311 [0.100]**
NPL dummy	1.897 [0.107]**	1.783 [0.150]**	1.873 [0.107]**	1.762 [0.150]**	1.868 [0.107]**
Public dummy	1.414 [0.332]**	0.807 [0.725]	0.501 [0.345]	0.471 [0.743]	0.499 [0.345]
With Author *NPL		0.222 [0.205]		0.228 [0.205]	
With Author*Public		0.759 [0.810]		-0.045 [0.836]	
With Crossover Inventor			1.455 [0.153]**	0.729 [0.292]*	1.661 [0.201]**
With Author*Crossover				0.991 [0.340]**	
After publication					0.285 [0.111]*
After pub*Crossover					-0.470 [0.284]
Constant	-1.663 [1.052]	-1.597 [1.054]	-1.737 [1.052]	-1.623 [1.053]	-1.919 [1.054]
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.13	0.13	0.13
# of observations	92,147	92,147	92,147	92,147	92,147

**: Statistically significant at 1% level, * Statistically significant at 5% level

Table 2-2: Regression results (Non-self Forward Citation, vs Same IPC patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	ALL
With paper author dummy	1.461 [0.088]**	1.262 [0.144]**	1.330 [0.089]**	1.047 [0.146]**	1.332 [0.089]**
NPL dummy	1.468 [0.095]**	1.310 [0.134]**	1.447 [0.095]**	1.291 [0.134]**	1.441 [0.095]**
Public dummy	1.812 [0.295]**	1.409 [0.644]*	0.988 [0.307]**	1.101 [0.661]	0.986 [0.307]**
With Author *NPL		0.308 [0.182]		0.314 [0.182]	
With Author*Public		0.501 [0.720]		-0.219 [0.744]	
With Crossover Inventor			1.313 [0.136]**	0.668 [0.259]*	1.547 [0.179]**
With Author*Crossover				0.883 [0.302]**	
After publication					0.283 [0.099]**
After pub*Crossover					-0.530 [0.253]*
Constant	-1.543 [0.936]	-1.456 [0.937]	-1.610 [0.935]	-1.480 [0.937]	-1.787 [0.938]
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.14	0.14	0.14
# of observations	92,147	92,147	92,147	92,147	92,147

**: Statistically significant at 1% level, * Statistically significant at 5% level

Table 2-3: Regression results (Generality Index, vs Same IPC Patents)

	(1)	(2)	(3)	(4)	(5)
Sample	ALL	ALL	ALL	ALL	ALL
With paper author dummy	0.022 [0.002]**	0.019 [0.003]**	0.019 [0.002]**	0.013 [0.003]**	0.019 [0.002]**
NPL dummy	0.018 [0.002]**	0.016 [0.003]**	0.018 [0.002]**	0.016 [0.003]**	0.018 [0.002]**
Public dummy	0.054 [0.007]**	0.002 [0.016]	0.036 [0.007]**	0.000 [0.016]	0.035 [0.007]**
With Author *NPL		0.004 [0.004]		0.004 [0.004]	
With Author*Public		0.062 [0.018]**		0.040 [0.018]*	
With Crossover Inventor			0.029 [0.003]**	0.006 [0.006]	0.039 [0.004]**
With Author*Crossover				0.030 [0.007]**	
After publication					0.004 [0.002]
After pub*Crossover					-0.024 [0.006]**
Constant	0.110 [0.029]**	0.111 [0.029]**	0.110 [0.029]**	0.113 [0.029]**	0.108 [0.029]**
Application year dummy	Yes	Yes	Yes	Yes	Yes
IPC subclass dummy	Yes	Yes	Yes	Yes	Yes
R2	0.13	0.13	0.13	0.13	0.13
# of observations	62,704	62,704	62,704	62,704	62,704

**: Statistically significant at 1% level, * Statistically significant at 5% level

Figure 1: AI and data driven innovation

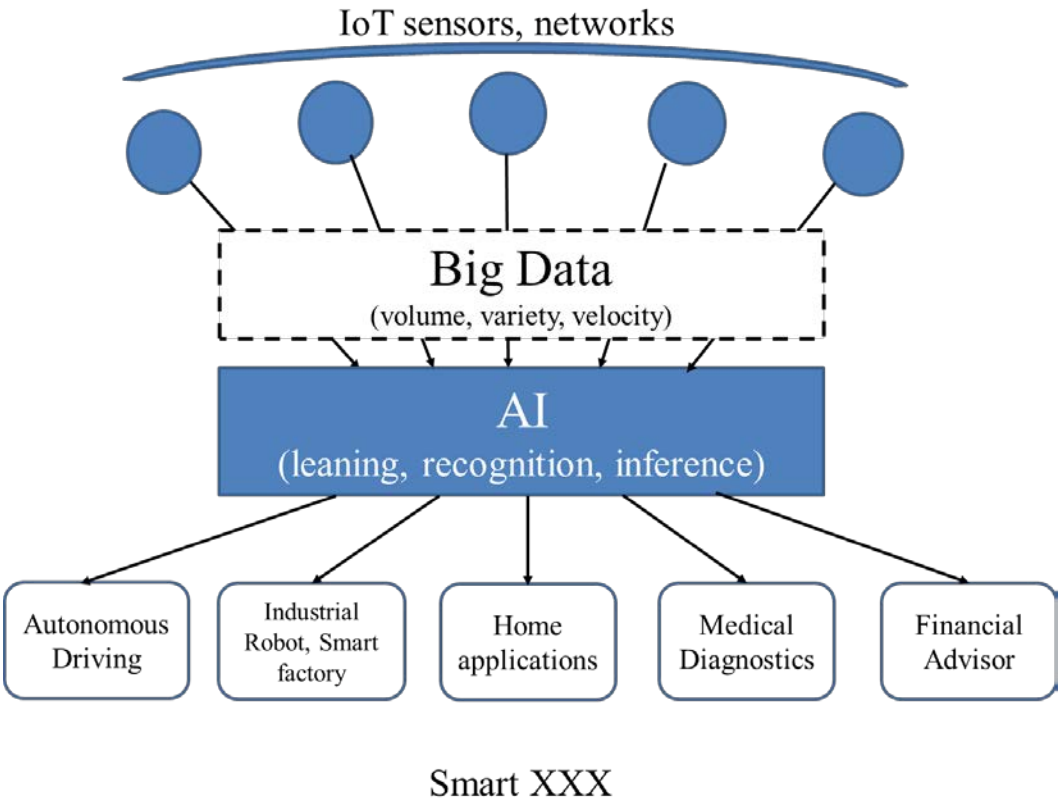

Figure 2: Share of AI papers and patents

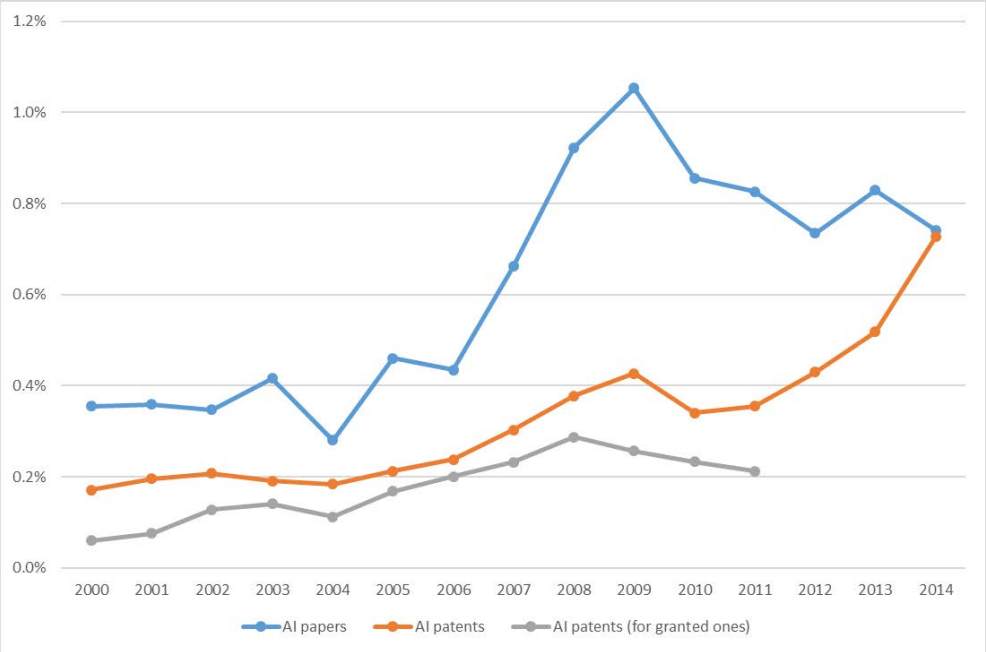

Figure 3: Share of private sector authors

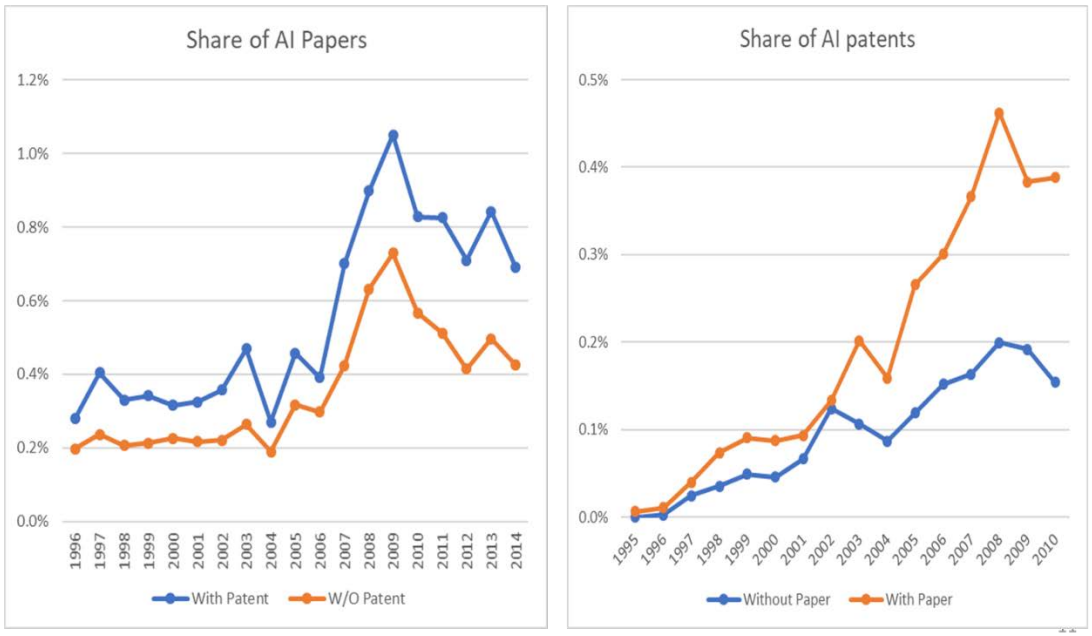

Figure 4: AI author and inventor patents by technology (except for computer technology)

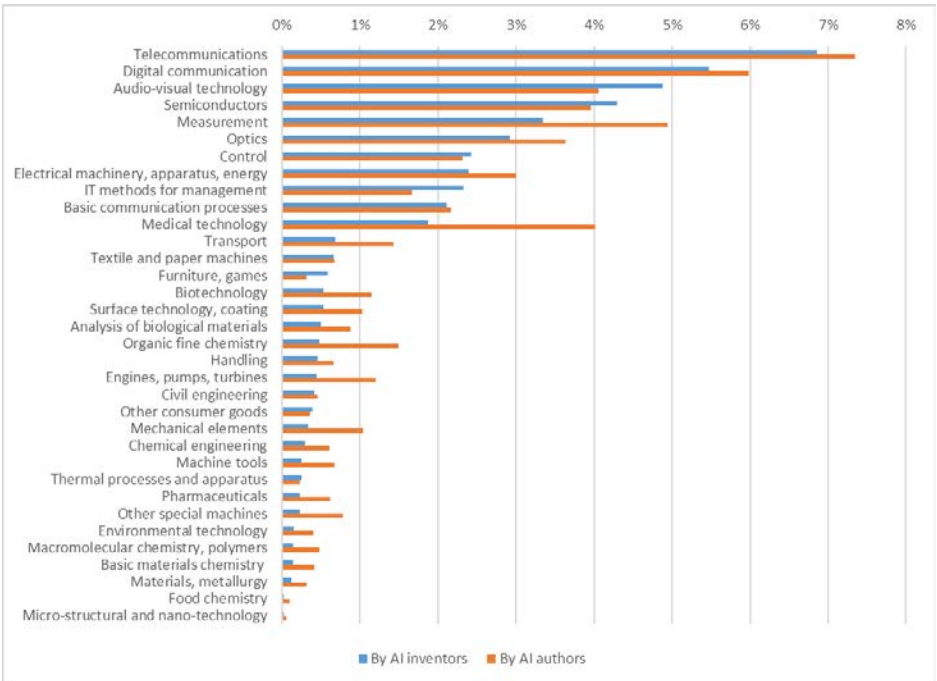

Figure 5: The share of patents with crossover inventor