Understanding Nobel Prize-winning articles: a bibliometric analysis

*Guoqiang Liang, Haiyan Hou, Peili Ren, Yi Bu, Xiangjie Kong and Zhigang Hu**

In the present study, we have collected all Nobel Prize-winning articles in the field of Physiology or Medicine from the Web of Science and highlighted the journal impact distribution of these articles. We then explored reference information to understand the articles referenced by the prize-winning papers. Results show that (1) the prize-winning papers cite a large number of journals which have relatively low impact factors and not all prize-winning papers were published in high-quality journals; (2) Method, such as refined experimental techniques, laboratory manuals, etc., is the most popular article type that has been cited; (3) The prize-winning papers, especially recently published ones, show an increasing trend to cite earlier published articles as references.

Keywords: Bibliometric analysis, normal and revolutionary science, Nobel prize-winning articles, reference information.

EVERY major scientific discovery has the potential to change the course of scientific development and even the world. It challenges acknowledged scientific convention and reorders old knowledge into a new paradigm, thus solving problems that stymied the scientific community[1]. Nobel Prize-winning articles (NPs) are such kind of work. For example, in 2006, Yamanaka *et al.* found that mature cells can be reprogrammed to become induced pluripotent stem cells (iPSCs) (note 1). Prior to this discovery, the most well-known iPSCs were embryonic stem cells. Researchers believed that embryonic stem cells could only be derived from the embryos, and this brings in an ethical issue because generation of embryonic stem cells involves destruction of the pre-implantation stage embryo. Since iPSCs can be derived from mature cells, Yamanaka *et al*. could successfully overcome the ethical issue and this technique is also used to generate transplants without the risk of immune rejection around the world (note 1). In 2012, Yamanaka won the Nobel Prize in Physiology or Medicine for the discovery of iPSCs.

Since NPs have deeply influenced the world and how we operate within it, it is essential to understand and derive insights from them to benefit research evaluation as well as resource allocation. In this study, we deal with the following: (1) Journal impact distribution of NPs and (2) 'Knowledge-base' (note 2) of NPs.

Guoqiang Liang, Haiyan Hou, Peili Ren and Zhigang Hu are in the WISE Lab, Dalian University of Technology, Dalian, Liaoning 116023, China; Xiangjie Kong is in School of Software, Dalian University of Technology, Dalian, Liaoning 116620, China; Guoqiang Liang, Yi Bu and Xiangjie Kong are also in the School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408, USA.
*For correspondence. (e-mail: huzhigang@dlut.edu.cn)

Related work

Theory and introduction of NPs

According to Kuhn[1], most scientific research can be categorized as 'normal science', which is viewed as 'development-by-accumulation', firmly based on past scientific achievements that are accepted by the established scientific community. However, normal science can be disrupted by revolutionary science (such as Copernicus' model of the universe, Newton's classical mechanics, Darwin's theory of evolution) which challenges commonly acknowledged principles and can successfully transform the established research paradigm. However, Kuhn did not provide specific criteria to distinguish between revolutionary and normal science. Casadevall and Fang[2] considered revolutionary science as a 'conceptual or technological breakthrough that leads to a dramatic advance in understanding', such as the emergence of new research. Charlton *et al.*[3] referred to NPs as revolutionary scientific research, and measuring competitive ability among different economic entities, they even confirmed that NPs may be the best indicator in judging the generation of new knowledge.

The Nobel Prize is a set of annual international awards, bestowed (in accordance to the will of the Swedish scientist Alfred Nobel) upon those who have made the 'greatest benefit to mankind' (note 3). The prizes in literature, peace, physiology or medicine, chemistry and physics were first awarded in 1901, and only to candidates who had made the most important discoveries or inventions in the above five domains. The Norwegian Nobel Committee awards one to three scientists per year for their outstanding work. Although the small number of Nobel

laureates annually indicates that most revolutionary achievements go unrecognized, only a small proportion of the awards are unjustified. The prize is the gold standard of scientific research in the fields where it is given[4,5]. Therefore, it makes sense to regard NPs as the most groundbreaking, transformative and revolutionary scientific research.

Research methods of NPs

Recently, Hu and co-workers[6,7] used the second generation of citations (citations of articles citing a work) to evaluate the most revolutionary work. They argued that NPs lead to profound transformative changes in current knowledge ecosystems and set NPs as an example to demonstrate that not all transformative works are highly cited. Mazloumian *et al.*[8] studied the citations of NPs based on 124 Nobel laureates awarded during 1990–2009, aiming to reveal how new ideas break through established paradigms. Tong and Ahlgren[9] employed citation analysis of NPs in chemistry to study the evolution of four themes in the field from an international cooperation point of view. Iwami *et al.*[10] took Yamanaka's NPs as a case study to argue that one can use the in-degree variation each year of an article to identify emerging leading papers, by constructing a direct citation network and computing the slope, area, height and time span of in-degree distribution of each article used. Wu *et al.*[11] used NPs to demonstrate how their 'disruptive index' can measure the influence of an article on a scientist by calculating the difference between the proportion of direction citations of article A and co-citation of article A with its citations.

While insightful, these previous methods have been unfortunately prone to a few biases. Time lag is inevitable when analysing citation information, which means that we have to wait five or ten years or even longer after the publication of an article to measure its quality. Moreover, just drawing information from the citation patterns of one or several NPs and then detecting 'emerging', 'leading' 'ground-breaking', or 'disruptive' papers is not objective, because the focus is only on a limited selection of NPs. Besides, there are different citation patterns after the publication of scientific articles. Finally, some quality papers have not received any citations for a long time after publication[12–14].

Knowledge base

It is commonly accepted that original ideas seldom come entirely 'out of the blue', they are often embodied in existing knowledge[15], like Newton's metaphor: 'if I have seen further, it is by standing on the shoulders of giants'. Our highly advanced modern society is largely shaped by the combination of existing science, technologies, values and concepts, culture and economics. We can easily capture 'footprints' of current achievements from existing knowledge. For example, synthetic biology is a rapidly emerging interdisciplinary branch of biology and engineering in the 21st century, which has the potential to fabricate practical organisms to recognize and destroy tumours, extend the organism's behaviour[16,17]. It represents the cutting-edge of biomedical field, which combines disciplines from biotechnology, control engineering, molecular engineering, biophysics, computer engineering, etc. The influence of past literature on recent research is manifested in citations[18]. Thus, we can capture footprints of knowledge flow and obtain insights into the history of science through analysis of these references[19,20].

In this study, we collected all NPs of physiology or medicine during 1901–2017 from the Web of Science (WoS), acquiring basic information as well as 'knowledge base' of NPs. We begin this article by introducing our data collection and filtering methods. Next, we present our results in a statistical, visual and textual manner. Finally, we interpret the implications as well as limitations of the study.

Data and methods

NPs retrieval

The data collecting process of NPs includes three steps: First, to find a Nobel laureate's corresponding name in WoS. A total of 108 prizes have been bestowed to 214 laureates in physiology or medicine starting with the first awards in 1901 (note 4).

Yet, queries using author full name were returned with no publications because not all Nobel laureates' full names are embedded in WoS. However, using the author name field allows us to target all Nobel laureates' publications. If the author's institution, research topic and name were the same as those provided by the Nobel Prize official website, Wikipedia and Google Scholar, we considered him/her to be a Nobel laureate. Secondly, we determined the specific research focus for which the Nobel Prize was awarded. We explored the Nobel Prize official website in order to select a Nobel laureate's research focus associated with the Prize. For example, when Henrik Dam was awarded the 1943 Nobel Prize in Physiology or Medicine, the committee declared that it was 'for his discoveries of vitamin K'. We then used 'vitamin K' as that Nobel laureate's research focus. In some cases, it was not easy to select the research focus based on the official website; in such circumstances we sought clues from their Wikipedia biography. For example, Karl Landsteiner won the 1930 Nobel prize for 'his discovery of human blood groups'. We then used 'blood', 'antigens', 'serum' and 'serological' as his research focus, based on information provided by Wikipedia. We measured the quality of keywords selected by comparing

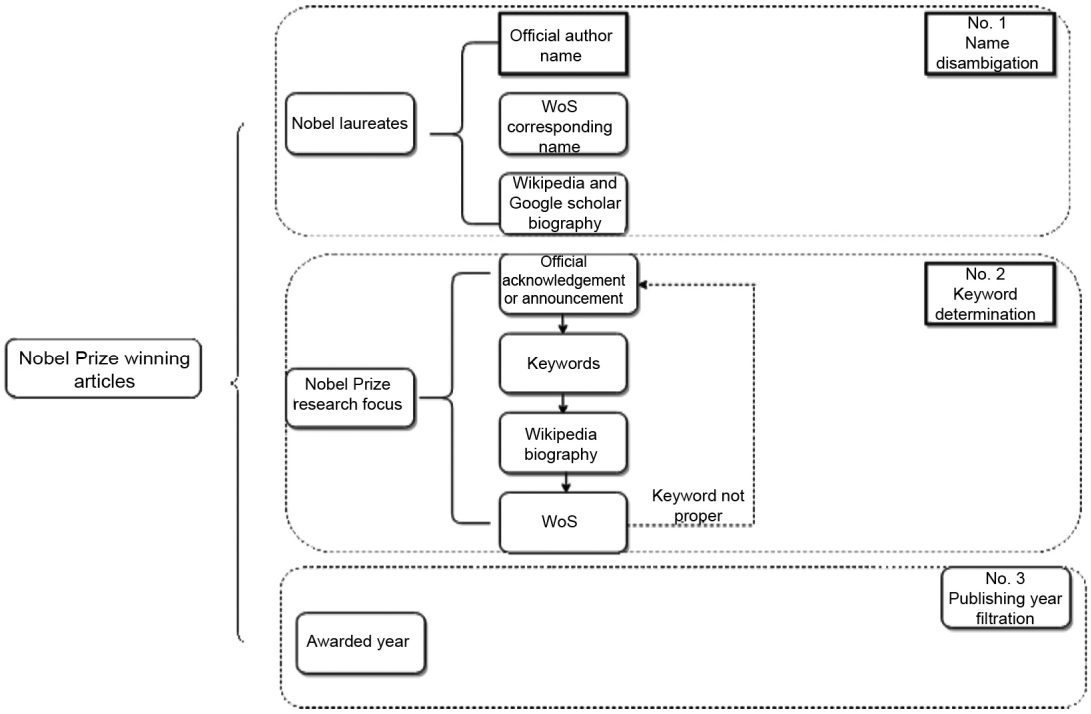

Figure 1. Identification procedures of Nobel Prize-winning articles.

them to results retrieved from WoS. Finally, we limited our definition of NPs to those published before the year Nobel laureates were awarded. This ensures that the work in our dataset is actually the body of research for which they were awarded the prize. Figure 1 shows the procedure used to retrieve NPs from WoS, resulting in a total of 7705 journal research articles.

Processing of references in NPs

Considering that the same references may have different reference formats in some cases, variants of the same references were eliminated. In the first step of elimination, variants within the same publication year were identified based on last name of author and source title. We then computed Levenshtein similarity after determining the pairwise similarity of variants. We considered two variants to be the same if their similarity value is above 0.75 (ref. 18). The Levenshtein similarity between two strings $s1$ and $s2$ is defined as

$$\mathrm{sim}(s1, s2) = 1 - \frac{\mathrm{LD}(s1, s2)}{\max(|s1|, |s2|)},$$

where LD ($s1$, $s2$) is the Levenshtein distance which is defined as the minimal number of single-character edit operations required to transform string $s1$ into $s2$. $|s|$ denotes the length of a string s. The Levenshtein similarity is between 0 and 1, where 0 indicates two totally different strings, whereas 1 indicates two identical strings. For more details, readers can visit https://github.com/Simmetrics/simmetrics (see also refs 21, 22). After applying this method, we determined that NPs had cited a total of 217,281 distinct references.

Estimation of citation time lag for NPs

A time-lag value was assigned to each article based on the median and mean of each time-lag vector. For example, we computed citation time-lag value of each NP in the same year and formed a citation time-lag vector in each year. We then selected a median and mean value to represent the time-lag for each year. The citation time lag (Ti) was computed by the publication year of an NP (PYi) minus the publication year of the references (PYij) cited by the NP. Where i is the ith article of NPs, j the jth reference cited by i and PY is publication year.

Results

Journal impact distribution of NPs

According to Kuhn[23], in general, there are two scientific strategies – tradition and innovation. In 1959, he introduced the notion of essential tension, arguing that once a certain scientific tradition becomes established, work within that field becomes more productive than when the field is first emerging, a strategy he called innovation.

Figure 2. Top journals with the most NPs. (The link represents impact factor of journals. The first column represents journal quartile based on *Journal Citation Reports*. The next column shows the ranking of journals based on the number of publications. The third column provides journal abbreviations and last column shows the total number of articles published.)

Foster *et al.*[24] divided the above two strategies into five: jump, new bridge, new consolidation, repeat bridge and repeat consolidation. The first three strategies correlated to varying degrees of innovation. Highly innovative research has high risk, because it may be rejected in the article review process for lack of common paradigm. NPs are risk-taking strategies to some extent. So where will these innovative achievements be published?

The results showed that 7705 NPs were published in 838 publications. We regard journals which published at least 51 articles as top journals according to their power law distribution. Figure 2 shows not all NPs were published in high quality journals; there are still some NPs published in Q4, Q3 journal quartile based on the 2016 *Journal Citation Reports*.

Knowledge base of NPs

Do high impact journals contribute the most knowledge to NPs? We analysed 207,636 source articles cited by NPs, which accounted for 95.6% of the total source referenced once we removed those without source titles. These cited articles originated in 11,493 titled sources, including journals, conference proceedings, patents and unpublished data. Also, 285 source titles occurred 68 times or more which accounted for 80.07% of the total source titles. Our analysis focused on those source titles which were journals, resulting in 242 journals in total. Results showed that NPs tend to cite articles published in very high impact journals (9 journals with an impact factor more than 30 contributed 17.35% knowledge to NPs) and those with average impact levels (208 journals with an impact factor below 14 contributed 55.6% the knowledge base of NPs) before being normalized by journal frequency (Figure 3). The Pearson correlation between journal impact factor and journal occurrence frequency was 0.258 (the 95% confidence interval between 0.050 and 0.457), which indicates there is a weak correlation that articles published in high impact journals contribute the most knowledge to NPs.

Which type of articles do NPs cite the most? We selected references which have been cited 79 times or more as the top references, which accounts for 1% of all references, after sorting according to correlation of times appeared and number of references (Figure 4 *a*). We obtained the citations of these top references using Google Scholar on 14 March 2018, and then read the abstract of each article to determine its type. We found that there were three types of articles (review, research and method). Method such as, refined experimental techniques, laboratory manuals, etc. is the most popular article type that has been cited (Figure 4 *b*).

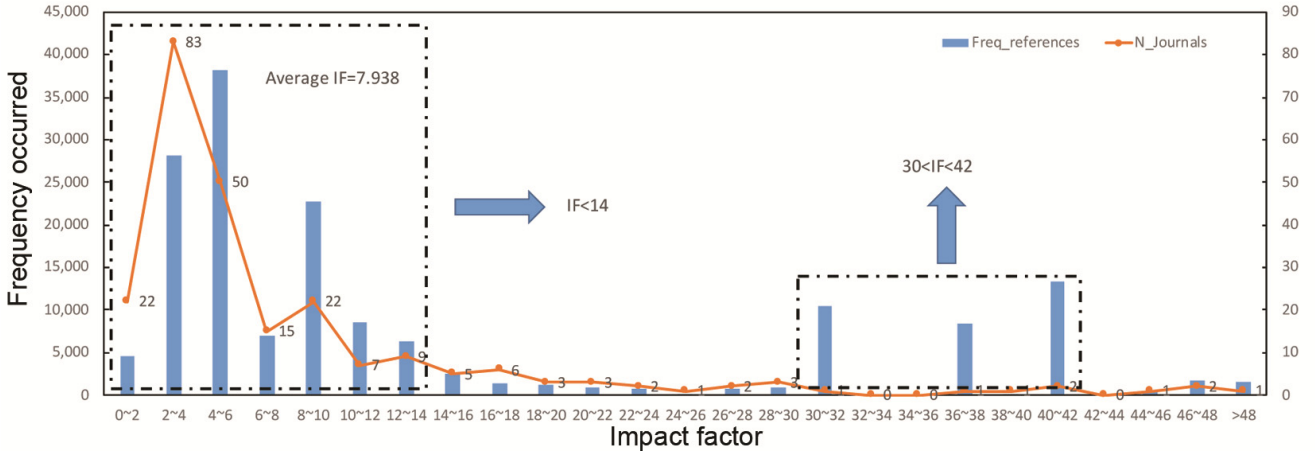

Figure 3. Distribution of journal impact factor of articles referenced by NPs.

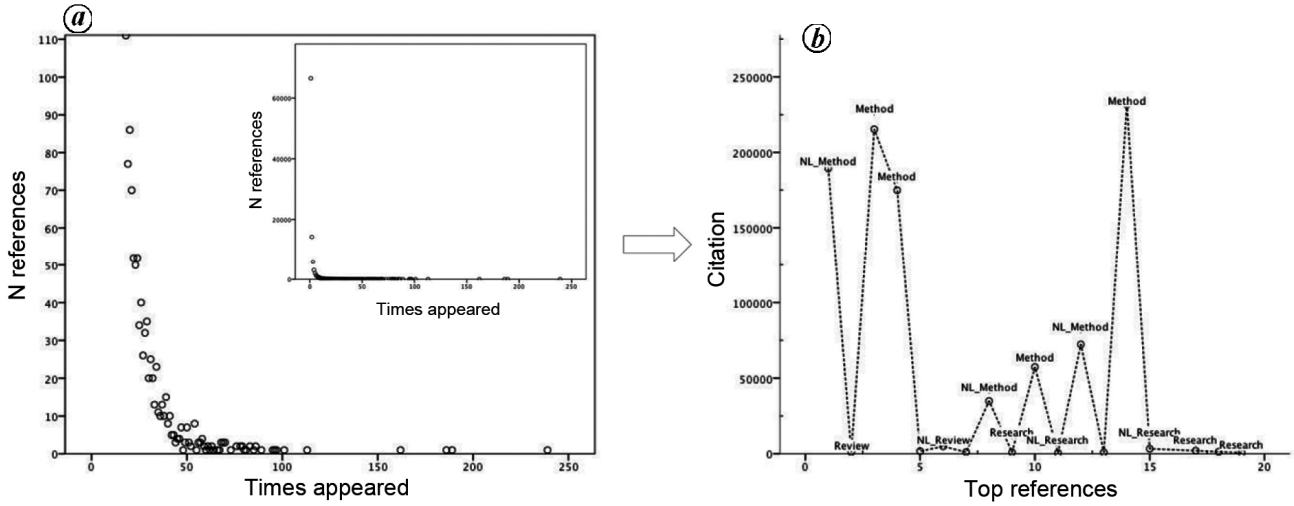

Figure 4. Distribution of (***a***) times appeared and number of references and (***b***) article type of top references.

Do NPs focus on recent achievements? Garfield[25] presented 15 reasons for reference citations. In recent years, researchers have found that citing behaviours are not only motivated by acknowledgement of prior scientists, but also by non-scientific factors, such as social or random factors[26]. Regardless of the motivation for which an author cites a reference, we cannot deny the fact that researchers focus more attention on recent achievements if they cite more recently published articles, which could also reflect the rapid development of a research area.

We analysed 213,956 source articles (98.47% in total) after excluding those without a publication year or unusual cases in which the publication year of the article cited was later than that of the citing article. We computed the time lag between references and citing article of NPs and formed a citation time-lag vector corresponding to each year. Then, we used median and mean to estimate the yearly distribution of time lag. The average time lag of all NPs was 6.15, which indicates that NPs tend to cite articles published 6.15 years ago on an average.

Figure 5 *a* illustrates the fluctuation of time lag during 1900–2017. According to the median, articles published during 1900–1901, 1915, 1918–1919, 1922, 1949, 2013–2017 are peaks in the time lag. On further inspection, we found an article titled 'Experiments on the value of vascular and visceral factors for the genesis of emotion' published in 1900, which cites two articles published in 1671 and 1677. The article 'Growth and regeneration in *Planaria lugubris*' published in 1901, cites five references published before 1833. The article 'The change in refractive power of the human eye in dim and bright light' published in 1947, cites a reference published in 1730. An article published in 2015 cites a reference published in 1891. Mitigating the effects of such outliers is the advantage of using the median of citation time lag. It is apparent that NPs focus more on earlier achievements, especially after 1980.

The price index computes the proportion of references within five years cited by an article; it is an indicator to measure literature obsolescence as well as to distinguish

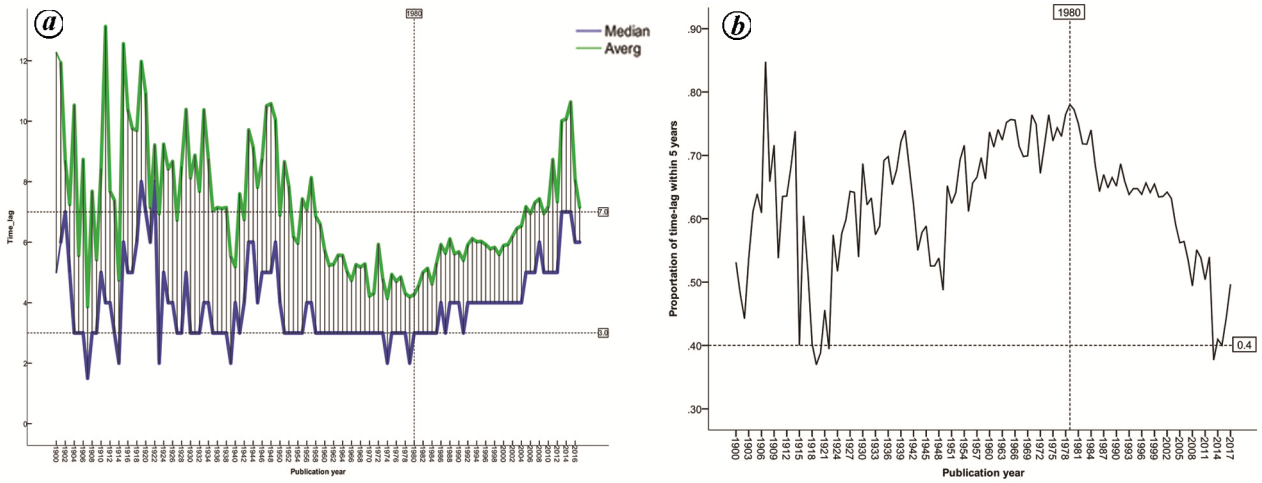

Figure 5. Time lag and rate of references within five years of NPs.

hard or soft science[27]. We computed the proportion of time lag equal to and less than 5 within our dataset, in order to find the obsolescence of articles within the Physiology or Medicine domain. Results showed that the proportion of references within 5 years cited by NPs decreased after 1980 (Figure 5). Theoretically, this may due to improved accessibility to older publications, and researchers in the domain of Physiology or Medicine attach high importance to former achievements. This highlights that while citing recently published references can reflect an author's attention to the latest developments, researchers should also pay attention to earlier achievements, at least in the area Physiology or Medicine, to improve the quality of an article.

Discussion and conclusion

Nobel Prizes are the most prestigious achievements of our time. Prizes in Physiology or Medicine are typically awarded for discoveries which have changed the direction of science[3]. Scientometric studies have designed or utilized various indicators to measure the quality of articles, to find key traits of revolutionary papers. This is essential for scientific policy as well as improving the work of scientists, but the more fundamental work is to understand the features of these articles. While there is no gold standard for revolutionary papers, in the present study we have restricted all the NPs in the Physiology or Medicine domain and considered them as verifiably revolutionary papers in order to find some patterns.

We report the following features of NPs: (1) They cite a large number of journals with relatively low impact factors, and not all NPs have been published in high-quality journals. (2) Methods, such as refined experimental techniques, laboratory manuals, etc. is the most popular article type that has been cited. (3) In recent years, the citation time lag of NPs has been increasing, which illustrates that recently awarded Nobel laureates focus more attention on source articles published in an earlier period.

This study is a fundamental work on NPs and there are some limitations. We collected data from the WoS platform, which means articles not recorded in WoS will not exist in our dataset, also papers without keywords or abstracts were hard to target. Additionally, there might be a potential data loss for NPs published in the early years due to the limited time span of WoS. For example, Emil Adolf von Behring was awarded the first Nobel Prize in 1901 for his work on serum therapy, but we could not obtain his articles published before 1900.

Further improvements need to be made: (1) Obtain a control group of non-NPs to find the unique feature of NPs. (2) Determine the driving force of different citation patterns of NPs as presented in Figure 3. (3) In-depth study of the relationship between knowledge recency and innovation based on NPs.

Notes

1. https://en.wikipedia.org/wiki/Induced_pluripotent_stem_cell
2. 'knowledge base', by which we mean the sources references in citations
3. https://www.nobelprize.org/alfred_nobel/will/will-full.html
4. https://www.nobelprize.org/nobel_prizes/medicine/laureates/index.html

1. Kuhn, T. S., Normal science as puzzle-solving. In *The Structure of Scientific Revolutions* (ed. Neurath, O.), The University of Chicago Press, Chicago, USA, 1963, pp. 35–43.
2. Casadevall, A. and Fang, F. C., Revolutionary science. *mBio.*, 2016, **7**, e00158-16.
3. Charlton, B. G., Measuring revolutionary biomedical science 1992–2006 using Nobel Prizes, Lasker (clinical medicine) awards and Gairdner awards (NLG metric). *Med. Hypotheses*, 2017, **69**, 1–5.

4. Rodríguez-Navarro, A., Measuring research excellence. *J. Doc*., 2011, **67**, 582–600.
5. Shelton, R. D. and Holdridge, G. M., The US–EU race for leadership of science and technology: qualitative and quantitative indicators. *Scientometrics*, 2004, **60**, 353–363.
6. Hu, X. and Luo, J., A warning for Chinese academic evaluation systems: short-term bibliometric measures misjudge the value of pioneering contributions. *J. Zhejiang Univ. Sci. B*, 2018, **19**, 1–5.
7. Hu, X. and Rousseau, R., Nobel Prize winners 2016: igniting or sparking foundational publications? *Scientometrics*, 2017, **110**, 1053–1063.
8. Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S. and Fortunato, S., How citation boosts promote scientific paradigm shifts and Nobel Prizes. *PLoS ONE*, 2011, **6**, 1–6.
9. Tong, S. and Ahlgren, P., Evolution of three Nobel Prize themes and a Nobel snub theme in chemistry: a bibliometric study with focus on international collaboration. *Scientometrics*, 2017, **112**, 75–90.
10. Iwami, S., Mori, J., Sakata, I. and Kajikawa, Y., Detection method of emerging leading papers using time transition. *Scientometrics*, 2014, **101**, 1515–1533.
11. Wu, L., Wang, D. and Evans, J. A., Large teams have developed science and technology; small teams have disrupted it. 2017; arXiv:1709.02445.
12. van Raan, A. F. J., Sleeping beauties in science. *Scientometrics*, 2004, **59**, 467–472.
13. Hu, Z. and Wu, Y., A probe into causes of non-citation based on survey data. *Soc. Sci. Inf*., 2018, **57**, 139–151.
14. Hu, Z. and Wu, Y. S., Regularity in the time-dependent distribution of the percentage of never-cited papers: an empirical pilot study based on the six journals. *J. Informetr*., 2014, **8**, 136–146.
15. Uzzi, B., Mukherjee, S., Stringer, M. and Jones, B., Atypical combinations and scientific impact. *Science*, 2013, **342**, 468–472.
16. Andrianantoandro, E., Basu, S., Karig, D. K. and Weiss, R., Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol*., 2006, **2**, 1–14.
17. Purnick, P. E. M. and Weiss, R., The second wave of synthetic biology: from modules to systems. *Nature Rev. Mol. Cell Biol*., 2009, **10**, 410–422.
18. Thor, A., Marx, W., Leydesdorff, L. and Bornmann, L., Introducing CitedReferencesExplorer (CRExplorer): a program for reference publication year spectroscopy with cited references standardization. *J. Informetr*., 2016, **10**, 503–515.
19. Garfield, E., Citation indexes for science. *Science*, 1955, **122**, 108–111.
20. Garfield, E., Historiographic mapping of knowledge domains literature. *J. Inf. Sci*., 2004, **30**, 119–145.
21. Thor, A., Bornmann, L., Marx, W. and Mutz, R., Identifying single influential publications in a research field: New analysis opportunities of the CRExplorer. *Scientometrics*, 2018, **116**, 591–608.
22. Wasi, N. and Flaaen, A., Record linkage using Stata: preprocessing, linking, and reviewing utilities. *Stata J*., 2015, **15**, 672–697.
23. Kuhn, T. S., The essential tension: tradition and innovation in scientific research? In *The Essential Tension*, The University of Chicago Press, Chicago, USA, 1976, pp. 225–240.
24. Foster, J. G., Rzhetsky, A. and Evans, J. A., Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev*., 2015, **80**, 875–908.
25. Garfield, E., Can citation indexing be automated? In *Statistical Association Methods for Mechanized Documentation*, National Bureau of Standards, Washington, USA, 1965.
26. Bornmann, L. and Daniel, H. D., What do citation counts measure? a review of studies on citing behavior. *J. Doc*., 2008, **64**, 45–80.
27. Price, D. J. D. S., Citation measures of hard science, soft science, technology, and nonscience. In *Little Science, Big Science…and Beyond*, Columbia University Press, New York, USA, 1986, pp. 155–180.

ACKNOWLEDGEMENT. This work is supported by the National Social Science Foundation of China under Grant No. 14BTQ030 and the China Scholarship Council.

Received 27 June 2018; revised accepted 18 October 2018

doi: 10.18520/cs/v116/i3/379-385