



A typology of scientific breakthroughs

Mignon Wuestman , Jarno Hoekman , and Koen Frenken 

Innovation Studies Group, Copernicus Institute of Sustainable Development,
Princetonlaan 8a, 3584 CB, Utrecht University, The Netherlands

an open access  journal



Citation: Wuestman, M., Hoekman, J., & Frenken, K. (2020). A typology of scientific breakthroughs. *Quantitative Science Studies*, 1(3), 1203–1222. https://doi.org/10.1162/qss_a_00079

DOI:
https://doi.org/10.1162/qss_a_00079

Received: 8 April 2020
Accepted: 2 June 2020

Corresponding Author:
Jarno Hoekman
j.hoekman@uu.nl

Handling Editor:
Ludo Waltman

Copyright: © 2020 Mignon Wuestman, Jarno Hoekman, and Koen Frenken. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



Keywords: breakthroughs, citation impact, scientific discovery, science of science, scientometrics, societal impact

ABSTRACT

Scientific breakthroughs are commonly understood as discoveries that transform the knowledge frontier and have a major impact on science, technology, and society. Prior literature studying breakthroughs generally treats them as a homogeneous group in attempts to identify supportive conditions for their occurrence. In this paper, we argue that there are different types of scientific breakthroughs, which differ in their disciplinary occurrence and are associated with different considerations of use and citation impact patterns. We develop a typology of scientific breakthroughs based on three binary dimensions of scientific discoveries and use this typology to analyze qualitatively the content of 335 scientific articles that report on breakthroughs. For each dimension, we test associations with scientific disciplines, reported use considerations, and scientific impact. We find that most scientific breakthroughs are driven by a question and in line with literature, and that paradigm shifting discoveries are rare. Regarding the scientific impact of breakthrough as measured by citations, we find that an article that answers an unanswered question receives more citations compared to articles that were not motivated by an unanswered question. We conclude that earlier research in which breakthroughs were operationalized as highly cited scientific articles may thus be biased against the latter.

1. INTRODUCTION

The evolution of scientific knowledge is commonly understood as an alternating process between short periods of breakthrough discoveries and long periods in which these breakthroughs are further refined and elaborated (Kuhn, 1962; Toulmin, 1967). Periods of breakthrough discoveries transform the knowledge frontier, while periods of refinement and elaboration allow for the major scientific, technological, and societal contributions of those breakthroughs to materialize (Evans, 2016; Hilgard & Jamieson, 2017; Winnink & Tijssen, 2014). However, while the periods of refinement and elaborations are well understood (e.g., Boyd & Richerson, 1985; Cavalli-Sforza & Feldman, 1981; Nelson & Winter, 1982), the characteristics of discoveries that transform the knowledge frontier remain unclear (Marx & Bornmann, 2013).

In recent years, several authors have suggested supportive conditions for scientific breakthroughs, such as cognitively diverse teams (Hage & Mote, 2010; Hinrichs, Seager, et al., 2017; Wu, Wang, & Evans, 2019), combinations of highly conventional and highly novel knowledge (Mukherjee, Romero, et al., 2017; Schilling & Green, 2011) and psychological characteristics of scientists, such as stubbornness and tenacity (Grumet, 2008). Generally, these studies understand scientific breakthroughs as being codified in highly cited publications, typically operationalized as the top 5%, 1%, or 0.1% highly cited articles in a field (Ponomarev, Williams, et al., 2014; Uzzi, Mukherjee, et al., 2013; Zeng et al., 2017). These studies thus treat scientific breakthroughs as a

group characterized by high impact and implicitly assume that all scientific breakthroughs are highly cited articles, and conversely that all highly cited articles are scientific breakthroughs. A systematic effort to identify different types of breakthroughs has so far not been made. This is, however, useful, as it might be the case that breakthrough types occur under different circumstances and differ in their citation impact.

Below, we develop a typology of scientific breakthroughs, and examine differences between kinds of breakthroughs in terms of their disciplinary occurrence, considerations of use and citation impact. We make use of the Charge-Chance-Challenge (Cha-Cha-Cha) theory of scientific discovery as described by Koshland (2007) to develop a typology of scientific breakthroughs. Rather than understanding scientific breakthroughs as either Charge, Chance or Challenge type, as Koshland does, we propose three discovery dimensions that underlie those three types to provide a better understanding of the varieties of scientific breakthroughs. Using the three dimensions, we test to what extent configurations of these three dimensions are observable in the scientific literature by qualitatively coding the full text of 335 articles that, according to experts, report on scientific breakthroughs. We then use these coded articles to explore how the different characteristics are distributed over scientific disciplines and vary in their considerations of use and citation impact. We are particularly interested in those aspects, because Koshland, in his paper, provides examples of the different types of discoveries that come primarily from physical sciences and life sciences. Furthermore, as it is known that scientific breakthroughs have transformational potential both within and beyond science (Winnink & Tijssen, 2014), we explore how the different typological dimensions relate to furthering fundamental understanding and considerations of use (Stokes, 1997). Finally, we model the effect of breakthrough characteristics on cumulative citation impact over 10 years by means of a set of regression models. We find that breakthroughs vary widely in their citation impact, and that there are telling differences in impact between different types of breakthroughs.

2. THE CHA-CHA-CHA THEORY OF SCIENTIFIC DISCOVERIES

Koshland's Cha-Cha-Cha theory of scientific discoveries was developed to aid in understanding the heterogeneous nature of major scientific advances, and to improve our understanding of the conditions under which scientific breakthroughs occur (Koshland, 2007). The theory is developed from the perspective that different field conditions lead up to different types of scientific discoveries. These field conditions relate to the perceived state of knowledge in a scientific field, which offers opportunities for scientists to make relevant scientific contributions (discoveries). For example, a scientific discovery may provide an answer to a long-standing question in a field. Alternatively, a breakthrough may be a serendipitous encounter with an important new piece of evidence, which may fit or question the existing theory or observations in a field. Scientists may also recognize a set of inconsistencies in the state-of-the-art literature in a field, which they aim to resolve. Koshland's Cha's summarize these different kinds of scientific discoveries in three types: Charge, Chance, and Challenge.

2.1. Charge, Chance and Challenge Type Discoveries

Koshland defines Charge type discoveries as discoveries that "solve problems that are quite obvious ... but in which the way to solve the problem is not so clear" (p. 761). In other words, Charge type discoveries resolve "known unknowns" (Logan, 2009). Koshland uses Isaac Newton's discovery of gravity as an example of a Charge type discovery, because "the movement of stars in the sky and the fall of an apple from a tree were apparent to everyone, but Isaac Newton came up with the

concept of gravity to explain it all” (p. 761). A recent example of a Charge type scientific breakthrough is “cloaking technology” (Leonhardt, 2006), an invisibility device that has been a long-standing dream of many scientists. While it had been proven that perfect invisibility is impossible due to the wave nature of light, there was reason to believe that “perfect invisibility within the accuracy of geometrical optics” was achievable (p. 1777). Leonhardt (2006) reports the formulation of a “general recipe” for the design of media that can achieve such invisibility with possibilities for practical demonstrations. This breakthrough has thus, at least in theory, solved a well-known problem in a way that had not been thought of before.

Koshland defines Chance type discoveries as “instances of a chance event that the ready mind recognizes as important and then explains to other scientists” (p. 761). For a Chance type discovery, the original contribution lies in recognizing the importance of an unexpected encounter or explaining the importance to other members of the scientific community. These encounters typically involve some kind of serendipity (Copeland, 2019; Koshland, 2007; Yaqub, 2018). Encounters may take the shape of accidental discoveries of fossils, ancient remains, and other natural phenomena, but also of unexpected outcomes of planned experiments, such as Alexander Fleming’s discovery of penicillin (Koshland, 2007). A more recent example of a Chance type discovery is reported in an article by Palmer, Barthelmy, et al. (2005), who report on neutron star SGR 1806-20 emitting a giant gamma-ray flare on 27 December 2004. Recognizing the importance of this event was crucial: The authors note that this flare was about a hundred times higher than the two giant flares observed from this neutron star earlier, whereas the energy of giant flares is usually a thousand times higher than that of a typical burst. Because of that difference, the authors further note that under different circumstances, this burst could have been interpreted as another type of burst. Instead, they suggest that the observed flare is of a newly discovered subclass.

As a third type of scientific breakthrough, Koshland defines Challenge type discoveries as “a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time” (p. 761). Koshland provides Einstein’s theory of special relativity as an example of a Challenge type discovery, as it provided a theory that explained anomalies with contemporary theories. Another example presented itself with the report on a draft sequence of the Neandertal genome (Green, Krause, et al., 2010). The authors emphasize in the introduction of the article that “substantial controversy surrounds the question of whether Neandertals interbred with anatomically modern humans” (p. 710). The challenged model was based on the idea that modern humans, after leaving Africa, completely replaced Neandertals without interbreeding. This theory was supported by evidence on morphological features and DNA of modern humans, although the evidence was considered to be inconclusive. The draft sequence of the Neandertal genome suggest that Europeans and Asians, but not Africans, have inherited genes from Neandertals—a finding that does not fit with the model. Instead, the authors put forward an alternative theory: Neandertals interbred with modern humans after they left Africa, but before they spread into Europe and Asia.

While Koshland’s categorization is intuitive, it has thus far not been used to systematically map scientific discoveries. One obstacle towards using Koshland’s theory to classify scientific breakthroughs holds that Koshland does not specify whether we should understand scientific discoveries as mutually exclusive (i.e., Chance, Charge, or Challenge alone), or as combinations of types. For example, a discovery can fit the description of both Chance and Challenge. Consider, as an example, an article published in 2000, reporting on the discovery of two early hominid skulls and tools at a site in the Republic of Georgia, which the authors interpret as evidence that “the initial hominid dispersal from Africa was driven not by technological innovation but more likely by biological and ecological parameters” (Gabunia, Vekua, et al., 2000,

p. 1025). This discovery fits the definition of a Chance type discovery because it involves a chance encounter that scientists recognized as important, but it also fits the definition of a Challenge type discovery because the authors interpret evidence that is incongruent with scientific theories of the time, and propose an alternative theory. Moreover, Koshland also does not specify whether the three types are meant to be exhaustive. It might be that some discoveries do not fit with the definition of any of Chance, Charge, or Challenge.

2.2. From Discovery Types to Discovery Dimensions

As we aim to characterize and compare scientific breakthroughs, we allow for the possibility that Koshland’s discovery types are neither exhaustive nor mutually exclusive. Rather, we assume that breakthroughs can be characterized on three binary discovery dimensions. For each of Koshland’s discovery types, the state of two of the three binary dimensions is fixed, while the state of the third dimension may vary (see also Table 1). For both states of each dimension, we provide examples of relevant scientific articles in Table A1. We summarize the dimensions as follows:

1. *The discovery is driven by a question, or by a research object*

First, we distinguish between discoveries that are question driven and discoveries that are research object driven. Whereas in the case of question-driven discoveries the area of ignorance and the line of enquiry in the field is well established and widely shared (“we know what we do not know”), discoveries driven by a research object are inverse question driven: The discovery precedes the formulation of the question (“we do not know what we do not know”) (Meyers, 2011). For example, archaeologists might discover ancient hominid remains in an unexpected location, which then raises questions about the distribution and social relation of hominids (Brunet, Guy, et al., 2002). The discovery of ancient remains thus drives the formulation of a question that was not asked before. Note that it may be the case that discoveries driven by a research object do actually provide answers to questions, but these questions were not the driver of the discovery.

Koshland’s Charge type can be characterized as question driven, referring to discoveries that solve long-standing problems. In our earlier example, the discovery team was able to design a theoretical cloaking device in response to the ambition of engineering invisibility. Chance and Challenge type discoveries are both research object driven: Chance type discoveries start from an encounter with a research object that awaits interpretation, and Challenge type discoveries start from the recognition of a research object that does not fit with existing theories. In our example of a Chance type discovery, this was the observation of a giant gamma-ray burst, and in our Challenge example, this was genetic

Table 1. Koshland’s types as configurations of discovery dimensions

Koshland’s types	Discovery dimensions		
	A: question-driven or research-object-driven	B: new or known question/research object	C: question/research object is against or in line with literature
Charge	question	either	in line with literature
Chance	research object	new	either
Challenge	research object	either	against literature

evidence against the model that claimed that modern humans replaced Neandertals without interbreeding.

2. *The discovery introduces a new question/research object, or contributes to a known question/research object*

As a second dimension, we distinguish between new and known questions and research objects. We understand questions and research objects that are known as questions and research objects that are documented in the scientific literature, and of which the scientists who made the discovery were aware. Conversely, new questions and research objects are those that are introduced by the scientists who made the discovery and are, therefore, themselves part of the discovery. With regard to the scientific impact of a discovery, this is a relevant distinction, as it indicates whether the discovery team can be credited for introducing the new question or research object, or only for resolving or contextualizing it. Koshland described this in terms of “uncoverers,” or scientific teams for whom uncovering the question or research object is (part of) their original contribution, and “discoverers,” or scientific teams that contribute to a question or research object uncovered by others (p. 761).

Koshland’s Chance type can clearly be characterized along this dimension. For Chance type discoveries, it is the “uncovering” that is critical, along with recognizing and interpreting the relevance of the uncovered research object. Without observing the giant gamma-ray burst, its discovery team would not have been able to report any discovery. In the case of Charge or Challenge discoveries, the question or research object can be either new or known. For Charge type discoveries, which answer “obvious” questions, the question may be a long-standing one that many others have tried and failed to solve, such as the ambition of invisibility or the puzzle of gravity, but it may also be a question that they raised themselves as an extension of existing literature. For example, the authors of an article that reports on the derivation of germ cells from stem cells argued that “because embryoid bodies sustain blood development, we reasoned that they might also support primordial germ cell formation,” thus raising the question of whether germ cells can indeed be made from such embryoid bodies (Geijsen, Horoschak, et al., 2004, p. 148). And, Challenge type discoveries can be a response to an accumulation of facts that the discovery team uncovers themselves, or that were already known in the literature before. Our example of a Challenge type discovery based on the draft sequence of the Neandertal genome includes both: It reports original evidence that counters the existing model, and describes pieces of evidence that were uncovered by others. Here, we are in agreement with Koshland (2007), who also argued that Challenge type discoveries can be accompanied with uncover or not. Following his wording, it is the discovery of a new explanation of facts that is critical for Challenge type discoveries, not the uncovering of the facts as such.

3. *The question or research object is against or in line with state-of-the-art literature*

Third, we distinguish between discoveries that go against state-of-the-art literature and discoveries that fit with or follow logically from existing literature. In other words, the discovery may have the potential to cause a paradigm shift, or it may fit within the current paradigm (Koshland, 2007; Kuhn, 1962).

Koshland’s Challenge type discoveries are driven by research objects that are incongruent with the current paradigm, and their interpretation thus calls for a paradigm shift. Challenge type discoveries can thus be characterized as “against state-of-the-art literature.” The article on the Neandertal genome, for example, reported existing evidence incongruent with the current paradigm, uncovered additional evidence, and offered an alternative model. Charge type discoveries, on the other hand, answer questions that have been part

of the existing literature or follow logically from it and, logically, cannot go against state-of-the-art literature. The discovery of a theoretical cloaking device was, indeed, in line with earlier ideas on the feasibility of such a device. Chance type discoveries may or may not be in line with state-of-the-art literature, depending on the interpretation of their discoverers. The article by Palmer et al. (2005) on an observed gamma-ray flare is an example of the former: the flare was interpreted as an additional category of flares. This interpretation offered an extension of the current model and did not require a paradigm shift. The discovery of hominid skulls and tools in the Republic of Georgia (Gabunia et al., 2000), by contrast, is an example of both a Chance type discovery and a Challenge type discovery, as the evidence is seen as incongruent with scientific theories of the time.

In summary, we can define Koshland's types as configurations of three binary dimensions, as summarized in Table 1. Following the table, Charge type discoveries are driven by a question, be it a new or known question, and are in line with existing literature. Chance type discoveries are driven by a new research object and may be in line with or against existing literature. Challenge type discoveries are driven by a new or existing research object, and go against existing literature.

It follows from Table 1 that Koshland's three discovery types are not exhaustive of the possible types that are analytically conceivable. Indeed, there is no reason to assume that discoveries will only meet the particular configurations of the dimensions that are consistent with Koshland's three types. Using the framework, we are able to characterize scientific breakthroughs in three binary dimensions, so that each scientific breakthrough is classified as one out of eight (2^3) possible types, rather than Koshland's three. The question, then, of which of the eight possible discovery types is most prevalent, is an empirical one.

3. RESEARCH DESIGN AND METHODS

3.1. Data Collection

To characterize different types of scientific breakthroughs, we make use of *Science's* annual announcement of the Breakthrough of the Year (BotY) (AAAS, 2018) between 1999 and 2012. Each year the magazine's scientific editors select "the most significant scientific discovery of the year" (AAAS, 2018) and its nine runners-up¹. The selected breakthroughs are described by the journal's reporters in the final issue of the year. These descriptions may refer to a single scientific breakthrough or to a multitude of breakthroughs that center on a common theme², and include a list of references to the original research described and other supportive material.

For this paper, we use the reference list of each BotY description to select research articles that report on the scientific breakthrough. We will refer to these articles as *breakthrough articles*. We use the following requirements in our selection of breakthrough articles from the BotY (and runners-up) reference lists: (a) Articles should be written in English³; (b) articles should be published in the same year as the year in which they were announced BotY or runner-up, with the exception of articles published in December the year before (as these were published after the

¹ In 2015, a People's Choice was introduced along with the editors' selection. However, as data from 2013 onwards are not collected, these People's Choice breakthroughs are not in the data. Before 2015, only *Science's* editors contributed to the selection process.

² For example, in the runner-up "Water, water everywhere" from 2000, one breakthrough is the discovery of an ocean on one of Jupiter's satellites, and another is proof that there has been water on Mars at some point in time.

³ No research articles written in another language were found.

BotY announcement of the previous year); (c) articles should be published in peer-reviewed academic journals⁴; (d) articles should report original results described in the BotY description: Review articles or articles that were included as further reading are omitted; (e) articles should have a DOI and be available on Web of Science (WoS); and (f) articles should not have been retracted afterwards⁵.

Although the announcement of BotYs began in 1996 (replacing the annual announcement of Molecule of the Year), data are collected from 1999 onwards, as no runners-up were announced in 1998. BotYs are collected until 2012, to allow the articles at least 6 years to receive citations after publication. This resulted in 335 scientific breakthrough articles, derived from 140 BotYs (14 years: one breakthrough and nine runners-up per year). Table 2 shows a summary of this selection process.

We used the DOI of each article to collect data from WoS: (a) publication date; (b) publication source; (c) citation report consisting of the number of citations per year for 10 years, or as many years as possible for articles published after 2008; (d) number of authors; and (e) a PDF of the article's full text (including abstract). These data were extracted from WoS in April 2018. The articles' full texts were used to code the breakthrough article in terms of the three discovery dimensions and its reported considerations of use.

We also collect data on the scientific discipline of each of our breakthrough articles based on the indexation of *Nature*⁶ (Springer Nature, n.d.). We distinguish between disciplines as listed by *Nature*: biological sciences⁷, business and commerce, environmental sciences, health sciences⁸, humanities, physical sciences, scientific community and society, and social sciences.⁹ Because many of the examples of Chance type discoveries supplied by Koshland are specifically from paleontological sciences and astronomical sciences, whereas examples from Charge and Challenge type discoveries are not (Koshland, 2007), we will further distinguish between paleontological sciences and other biological sciences and between astronomical sciences and other physical sciences.

3.2. Coding Discovery Dimensions and Reported Considerations of Use

We use directed content analysis (Hsieh & Shannon, 2005; Saldaña, 2015) to code each breakthrough article on each of the three binary discovery dimensions and on the considerations of use of the scientific breakthrough article. The result of this process will be used to assess descriptively, statistically, and visually differences in discipline, citation impact, and reported considerations of use between breakthrough articles by dimensions (see also Section 2.2).

To code articles on the three discovery dimensions, we use the text of the articles. For each article we search for key phrases indicative of each dimension. Examples of key phrases used can be found in Table A1, and were developed in three steps. First, two coders, K. S. and M. L. W.,

⁴ Note that, although BotYs are announced by *Science*, reference lists include articles published in other peer-reviewed journals.

⁵ As retraction of articles is essentially right-censored, because any article may be retracted in the future, there may be a bias against older articles. However, there are few retractions in the data.

⁶ For breakthrough articles that were not published by *Nature*, we identify the most relevant scientific discipline by determining the discipline of referenced articles published in *Science*.

⁷ Including anatomy, physiology, cell biology, biochemistry, biophysics, and paleontology (Springer Nature, n.d.).

⁸ Including aspects of health, disease and healthcare aiming to develop knowledge, interventions, and technology for use in healthcare (Springer Nature, n.d.).

⁹ However, in our data set, we only find articles from "biological sciences," "physical sciences," "environmental sciences," and "health sciences."

Table 2. Article selection process

Requirement met	# articles
Total listed references	895
Written in English	895
Published in announcement year	746
Published in peer-reviewed journal	691
Reports original results	367
Have DOI and in Web of Science	340
Have not been retracted	335

coded 14 breakthrough articles from 2014, which were not part of the data set of this paper, by highlighting phrases that signal the state of the three dimensions as defined in Section 2.2. During this stage it was found that while relevant phrases could be found throughout the whole text of the article, the articles' abstracts, first sentence, introductions and conclusions are the most informative with regard to the state of the discovery dimensions. Coding thus focused on these sections, or on the whole text if abstract, introduction, and conclusion were inconclusive. Second, coding differences between K. S. and M. L. W. were discussed until a consensus was reached on common coding practices. Third, the highlighted phrases of these 14 breakthrough articles were summarized into stylized phrases. Note that some key phrases serve as signals for more than one dimension. For example, the phrase "On [date] we have observed [...]" signals that the reported scientific breakthrough is research object driven rather than question driven, but also that this research object is new rather than known, because uncovering this research object is part of the breakthrough. Fourth, the 335 breakthrough articles in our data set were then independently coded by both coders. For our analyses, the dimension states *question-driven*, *new*, and *against literature* are coded as 1, and *research object-driven*, *known*, and *in line with literature* are coded as 0.

For the identification of reported considerations of use we follow the four quadrants proposed by Stokes (1997) when cross-tabulating two questions: (a) Does the article report applied considerations of use of the scientific breakthrough, or not?; and (b) Does the article report that the scientific breakthrough is part of a quest for fundamental understanding, or not? Articles that do not report applied considerations of use but do report contributions to fundamental understanding are considered basic research. Articles that only report applied considerations of use without contributing to fundamental understanding are applied research. Articles that report both applied and fundamental considerations of use are considered as use-inspired basic research, also known as "Pasteur's quadrant" (Stokes, 1997). Finally, articles may report neither applied nor fundamental considerations of use. For the development of key phrases that signal considerations of use, we followed the same procedure as for the development of key phrases for discovery dimensions, described above. Such phrases were found to be typically reported in the final paragraph(s) of the breakthrough articles. Key phrases, as well as examples of reported considerations of use, can be found in Table A2.

We present intercoder reliability for each of our coded dimensions in Table 3, where we report Cohen's kappa, which takes intercoder agreement by chance into account. We find that kappa values are sufficiently high (Cohen, 1960). In cases of disagreement, final codes are based on consensus between the two coders. Consensus was found for all articles in the data set, which implies that all 335 publications originally selected serve as empirical observations.

Table 3. Intercoder reliability

Dimension	Cohen's kappa
A: Question or research object	0.756***
B: New or known	0.768***
C: Against or in line with literature	0.700***
Applied	0.760***
Basic	0.508***

Note: * $p < .05$, ** $p < .005$, *** $p < .001$.

3.3. Analysis

We run χ^2 tests and Tukey's HSD post hoc tests (Tukey, 1949) to test whether discovery dimension states are associated with scientific disciplines and reported considerations of use. We also present bar charts to assess differences visually.

To test whether discovery dimensions affect cumulative citation patterns, we run a set of regressions with the number of cumulative citations as the dependent variable, and three dummies that represent the three binary discovery dimensions as the main independent variables. Ten regression models estimate cumulative citations from 1 to 10 years after publication. For this, negative binomial regression is appropriate, as our dependent variables reflect overdispersed count data (Cameron & Trivedi, 1998): If we were to use a Poisson regression rather than a negative binomial regression to model cumulative citations 10 years after publication using all our independent variables, the residual variance of 184.073 would exceed the 220 degrees of freedom. As each BotY description can contain references to multiple breakthrough articles, we cluster standard errors at the level of the BotY description. As using cumulative citations per year makes it difficult to identify differences in the number of citations per year, we rerun our models using number of citations per year for one to 10 years after publication rather than cumulative citations per year for 1 to 10 years after publication as dependent variables, presented in Figure A1 and Table A3.

Individual authors may each boost the cumulative citations to their own articles by bringing their work under the attention of others (Aksnes, 2003). Therefore, we include *number of authors* as a control variable. As this variable is heavily skewed, we use a log transformation for *number of authors*. In a second set of models, we further include dummy variables for *discipline*, as it is known that citations rates vary between disciplines, and we find that configurations of discovery dimensions are not randomly distributed over disciplines.

4. CHARACTERIZING BREAKTHROUGHS

4.1. Configurations of Discovery Dimensions and Associated Disciplines and Considerations of Use

In Table 4, we present the distribution of articles in our data set over the eight different configurations of discovery dimensions. We also compare our typology to Koshland's. We see that some combinations of characteristics are more common than others. Notably, the majority of our articles (77%) are what Koshland would describe as Charge type discoveries: driven by a known question that is in line with theory, irrespective of whether the research object is new or known. We also find that most articles can indeed be classified according to Koshland's typology: Only 43 articles (13%) do not fit with that typology, either because they have

Table 4. Configuration of discovery dimensions

	Discovery dimensions			N	%	Koshland's type
	A	B	C			
A0B0C0	research object	known	in line with literature	8	2.4	-
A0B0C1	research object	known	against literature	2	0.6	Challenge
A0B1C0	research object	new	in line with literature	31	9.3	Chance
A0B1C1	research object	new	against literature	5	1.5	Challenge + Chance
A1B0C0	question	known	in line with literature	166	49.6	Charge
A1B0C1	question	known	against literature	16	4.8	-
A1B1C0	question	new	in line with literature	93	27.8	Charge
A1B1C1	question	new	against literature	14	4.2	-

properties of more than one type (11%), or because they have properties of none (2%). In this sense, the original typology of Koshland can be regarded as useful.

Figure 1 presents bar charts of the discipline and reported considerations of use for each configuration of discovery dimensions, discussed in detail below. Most articles in our data set report considerations of use for fundamental understanding only (74%). Just a few are only applied (7%), while another 17% are classified as both fundamental and applied; 2% report neither.

A large share (47%) of the articles report on research on biological sciences (excluding palaeontology), while 26% report on research on physical sciences (excluding astronomy). Research in the field of health sciences and environmental sciences is less common.

4.1.1. Dimension A

The majority of articles in this study (86%) are question driven. We find that question driven discoveries are not randomly distributed across disciplines ($\chi^2 = 100$, $df = 5$, $p < .001$). Based

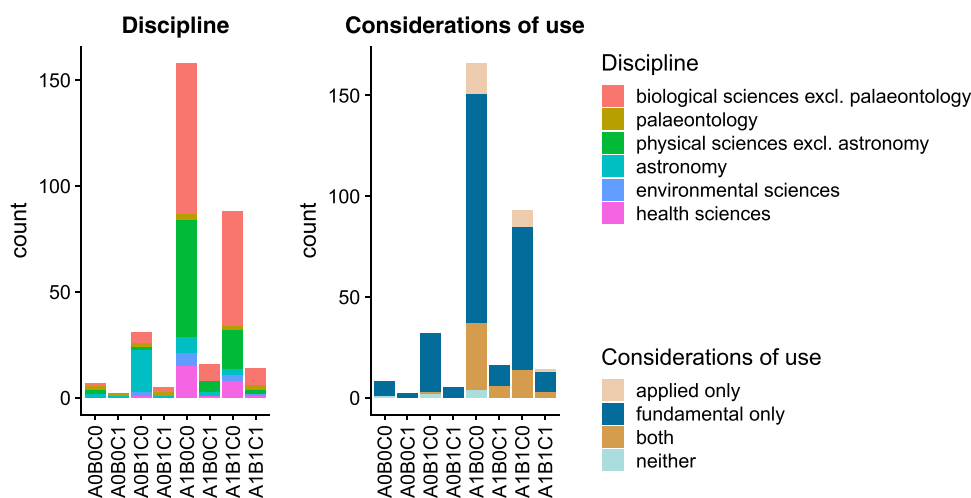


Figure 1. Bar chart of disciplines (left) and considerations of use (right) per configuration of discovery dimensions.

on our post hoc test, we find that being question driven is associated with health sciences, more than with other disciplines ($p < .05$). Indeed, almost all (96%) health sciences articles in our data set are question-driven. Conversely, being research object driven is associated with astronomy and paleontology more than with other disciplines ($p < .05$). This may be because disciplines such as paleontology and astronomy more often encounter unexpected physical research objects, for example from fossil records and satellite observations, respectively. Question-driven articles are also not randomly distributed across the four Stokes quadrants ($\chi^2 = 20$, $df = 3$, $p < .01$). Our analysis suggests that being question driven is associated with reporting only applied considerations of use and with reporting both applied and fundamental considerations of use, while research object driven breakthroughs are associated with reporting on neither ($p < .05$).

4.1.2. Dimension B

Articles with a new question or research object are slightly less common than articles with a known question or research object (43% versus 57% of articles). Most of these are articles with a new question rather than a new research object (74%). Of our total set of articles, only 10% report on a new research object that is in line with theory.

Our results indicate that this dimension and discipline are not independent ($\chi^2 = 20$, $df = 5$, $p < .001$). Specifically, we find that having a new question or research object is associated more with biological sciences (but not paleontology) and astronomy, while having a known question or research object is associated with physical sciences (but not astronomy) ($p < .05$). It is further worth noting that breakthroughs that are specifically driven by a new research object are primarily found in astronomy (see also Figure 1). We do not find a strong association between this discovery dimension and Stokes' four quadrants regarding considerations of use ($\chi^2 = 7$, $df = 3$, $p > .05$).

4.1.3. Dimension C

Breakthroughs that go against the state-of-the-art literature are uncommon (11%) and among them the large majority are question driven. We do not find a strong association between this discovery dimension and discipline ($\chi^2 = 9$, $df = 5$, $p > .05$). Our post hoc tests suggest that breakthroughs going against the literature are somewhat common in paleontological articles, while being in line with the literature is associated more with health sciences ($p < .1$) and physical sciences except astronomy ($p < .05$). In terms of reported considerations of use, we do not find significant evidence that being against state-of-the-art literature is associated with reported considerations of use ($\chi^2 = 5$, $df = 3$, $p > .1$).

4.2. Citation Impact and Discovery Dimensions

Table 5 shows descriptives of the cumulative number of citations within 10 years per discovery dimension state. On average, the articles in our data set collect 799 citations within the first 10 years after publication. However, with a median of 489 and an interquartile range of 625, this varies broadly: While the lowest decile of the articles in our data set have fewer than 108 citations, the highest decile has more than 1,571. Interestingly, one article did not receive any citation within 10 years¹⁰.

¹⁰ This article, which reports the first results from the Sudbury Neutrino Observatory (Helmer & SNO Collaboration, 2002). It may be that this article did not receive any citations because there were two other articles that report results from the same observatory (also included here). While all three articles were originally submitted in April 2002, the Helmer et al. paper was published in November, while the other two were published in June.

Table 5. Summary statistics of cumulative citations within 10 years per discovery dimension state

	Dimension	<i>N</i>	Mean	St. dev.	Min	1st quartile	Median	3rd quartile	Max
A1	question-driven	288	867.1	1,203.9	27	268	526	917	9,356
A0	research-object-driven	47	412.3	433.2	0	150	302	418	2,011
B1	new	144	740.8	1,094.8	0	240	444	738	9,356
B0	known	191	846.6	1,166.6	27	247	505	917	8,780
C1	against literature	37	769.8	962.0	27	219	519	898	4,117
C0	in line with literature	298	803.2	1,157.3	0	246	484	868	9,356
	total	335	799.2	1,134	0	245	489	870	9,356

To test if the discovery dimension can explain some of the variation in citation counts, we present the incidence rates of negative binomial regression models including control variables in Figure 2, with one regression for each of the 10 years. Incidence rates for the effect of being *question-driven*, driven by a *new* question or research object, or being *against literature* are presented relative to being *research object-driven*, driven by a *known* question or research object and being *in line with literature*, respectively.

Table 6 presents the regression coefficients of our models, where cumulative citations 1, 2, 5, and 10 years after publication are used as dependent variables. Models based on cumulative citations after 3, 4, 6, 7, 8, and 9 years were omitted from this table for readability reasons. Dummies for the three binary dimensions (with *question-driven*, *new*, and *against literature* coded as 1, and *research object-driven*, *known* and *in line with literature* coded as 0) and control variables for # *authors* (log) and *discipline*, with biological sciences as reference category, are included.

4.2.1. Dimension A

We find that being question driven has a positive effect on cumulative citations of scientific breakthrough articles. After 10 years, articles that are question driven are estimated to receive twice as

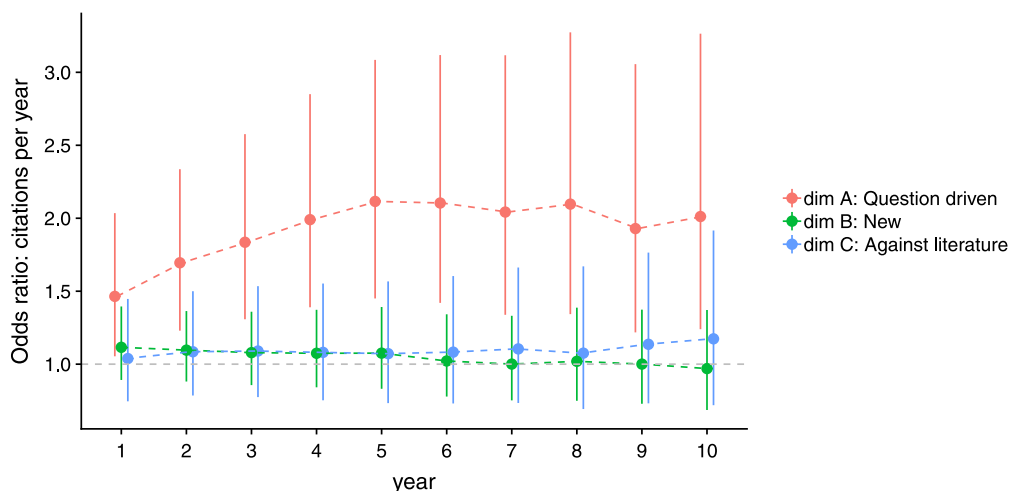


Figure 2. Incidence rate ratios based on negative binomial regression on cumulative citations after *N* years, after controlling for the log of the number of authors and discipline.

Table 6. Coefficients for negative binomial regression models

	cumu_1y	cumu_2y	cumu_5y	cumu_10y	cumu_1y	cumu_2y	cumu_5y	cumu_10y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
# authors (log)	0.360*** (0.120)	0.337*** (0.096)	0.300*** (0.086)	0.195 (0.143)	0.344*** (0.119)	0.319*** (0.100)	0.263*** (0.091)	0.176 (0.144)
Dimension A	0.570*** (0.303)	0.655*** (0.310)	0.827*** (0.422)	0.724* (0.586)	0.382* (0.254)	0.528** (0.278)	0.749*** (0.409)	0.699* (0.602)
Dimension B	0.174 (0.113)	0.135 (0.099)	0.110 (0.108)	0.009 (0.138)	0.110 (0.107)	0.092 (0.096)	0.073 (0.107)	-0.030 (0.148)
Dimension C	0.090 (0.171)	0.109 (0.177)	0.086 (0.203)	0.148 (0.309)	0.038 (0.165)	0.082 (0.196)	0.069 (0.232)	0.160 (0.361)
Constant	3.036*** (0.202)	3.749*** (0.195)	4.677*** (0.225)	5.576*** (0.304)	3.424*** (0.230)	4.036*** (0.224)	4.875*** (0.264)	5.752*** (0.355)
Discipline	excluded	excluded	excluded	excluded	included	included	included	included
Observations	335	335	335	244	335	335	335	244
LL	-1,840	-2,084	-2,393	-1,866	-1,758	-1,993	-2,290	-1,755
AIC	3,691	4,177	4,796	3,743	3,533	4,001	4,596	3,527

Note: *** $p < .001$, ** $p < .01$, * $p < .05$.

many citations as articles that are research object driven (Model (8): $IR = e^{0.699} = 2.011$). There is a smaller positive effect immediately after publication (Model (5): $IR = e^{0.382} = 1.465$), which increases over time and seems to stabilize after 4–5 years and decrease slightly after 8 years. Controlling for *discipline* in Models 5–8 slightly reduces the effect of being question driven, suggesting that part of the effect seen in Models 1–4 is, in fact, due to high citation rates of disciplines that are associated with being question driven. However, this does not alter our conclusion that being question driven has a positive effect on cumulative citations of scientific breakthrough articles.

4.2.2. Dimension B

We do not observe a significant association between this dimension and cumulative citations. Our coefficients suggest that there may be a small positive effect of being driven by a new question or research object on cumulative citations shortly after publication, which decreases in later years. This may be caused by the unexpectedness and novelty of the new question or new physical evidence introduced in the breakthrough article, and the sudden interest that this may spark. However, this is not a significant finding.

4.2.3. *Dimension C*

We do not find a significant association between articles going against the literature and cumulative citations. Upon visual inspection, there is some indication that breakthrough articles driven by a question or research object that is against state-of-the-art literature receive more citations in later years (year 9 and 10). The trend observed supports the idea that paradigm-shifting discoveries require more time to have an impact before they can be integrated in future knowledge development. However, the results are statistically insignificant.

The results of Models 9–16, where we use citations per year rather than cumulative citations per year as dependent variable, are presented in Figure A1 and Table A3. These results are in line with our earlier results. Again, we find that only the question-driven dimension significantly affects the number of citations received. We find that the difference in the number of citations per year between question-driven and research object-driven articles is biggest after 4–5 years.

5. DISCUSSION

In this paper, we have developed a typology of scientific breakthroughs and applied this typology to characterize a set of articles reporting on scientific breakthroughs. Using Koshland's Charge-Chance-Challenge theory of scientific discovery as a starting point, we propose that scientific breakthroughs can be characterized along three dimensions: (a) whether the discovery is question driven or research object driven; (b) whether the discovery contributes to a known question or research object or introduces a new one; and (c) whether the discovery is in line with, or against, state-of-the-art literature. We subsequently use the typology to characterize 335 breakthrough articles along the three dimensions and analyzed how breakthrough characteristics relate to scientific disciplines, citation impact, and considerations of use for fundamental understanding and application.

One of our main findings holds that the large majority of breakthrough discoveries can be classified as one of Koshland's discovery types within his Cha-Cha-Cha framework. However, we also observed that a small proportion of breakthroughs could not be characterized as any of Koshland's types, and some other articles fell into multiple Koshland types. Based on this finding we conclude that, rather than distinguishing between Charge, Chance, and Challenge types, breakthroughs can better be understood as being question driven or research object driven, introducing a new question/research object or a known question/research object, and having a contribution that is against or in line with state-of-the-art literature. We believe that our framework marks an improvement over the original Cha-Cha-Cha theory, as we have made the underlying dimensions explicit and orthogonal to one another, expanding the typology from 3 to $2^3 = 8$ types. Our framework, then, can be used in future research to further probe the antecedents and effects of scientific breakthroughs. It can equally be used to analyze differences between characteristics of breakthrough and nonbreakthrough discoveries. A logical extension of this paper is also to study whether the configurations of discovery dimensions discussed here are distributed differently over breakthroughs than over nonbreakthroughs, and to test whether the citation patterns we found are also observed for nonbreakthroughs.

Our main empirical finding holds that most scientific breakthroughs are driven by an already existing question and in line with the state-of-the-art literature. This finding broadens our view of science in that it questions the popular view of scientific breakthroughs as radical, paradigm-shifting discoveries (e.g., Evans, 2016; Ventegodt & Merrick, 2004). Rather, it suggests that the majority of scientific discoveries that are recognized as breakthroughs are better described as "normal science" (Kuhn, 1962).

Our analysis also shows that articles reporting on scientific breakthroughs vary considerably in their citation impact. In particular, breakthrough articles that were driven by a research object rather than a question receive far fewer citations. This finding has implications for the interpretation of earlier research on scientific breakthroughs. Previous research has mainly analyzed scientific breakthroughs based on citation impact, thereby considering breakthroughs as a homogeneous group of discoveries (Ponomarev et al., 2014; Uzzi et al., 2013; Zeng et al., 2017). In contrast, our findings suggest that earlier research aimed at identifying supportive conditions for scientific breakthroughs did not recognize the variety of breakthroughs and may have been biased against a minority of breakthroughs driven by research objects. Therefore, their findings may not be generalizable to all scientific breakthroughs. For literature on scientific breakthroughs, a next step is to identify how conditions such as team composition and sponsoring affect the occurrence of a variety of scientific breakthroughs, and in particular those breakthroughs that are research object-driven, as we have shown that these have been underrepresented in the literature thus far.

In this research, discoveries that were marked as scientific breakthroughs by *Science* have been leading. This operationalization of scientific breakthroughs has several implications for the generalizability of our findings. In the first place, our research only includes discoveries that are recognized as scientific breakthroughs within a year after publication. Discoveries that are recognized as such in a later phase may not have the same characteristics. For example, their citation impact may differ over time. An interesting avenue for future research would be to distinguish between discoveries that are received as breakthroughs shortly after publication and those that are recognized as breakthroughs later on. An interesting question then holds whether the relative prominence of the three dimensions introduced here differs between early and delayed recognition. In the second place, it is not unlikely that the nomination for BotY in itself affects the way a discovery is received. The increased visibility of the discovery may inspire others to refine the discovery in other research projects, and can lead to an increase in citations or even an increase in the likelihood of receiving a significant prize, such as a Nobel Prize or a Fields Medal. For future research, we encourage alternative approaches to identifying scientific breakthroughs that are more sensitive to delayed recognition and are not based on external assessments. One such approach has been developed by Small, Tseng, and Patek (2017), who identify and characterize biomedical discoveries based on automated text analysis of citing sentences and cocitation analysis.

Our analysis of breakthrough discoveries is further limited by what has been reported in the scientific articles. As such, we must limit ourselves to an analysis of the *reported* drive of the scientific discoveries observed, which may not be the same as the actual drive of the discovery. Indeed, authors may present the process of discovery as more linear and rational than it actually has been (Myers, 1985). Similarly, the authors' motivation to write and publish the article may be different from their motivation to start the reported research project. For example, their original line of enquiry may have resulted in a serendipitous finding that solves an unexpected problem in another line of enquiry (Yaqub, 2018), which might lead the authors to change their narrative as well.

Furthermore, our analysis is limited by the limited number of articles considered. With more observations, we could test differences in citation patterns of combinations of dimensions, rather than for single dimensional states. This may help us understand whether breakthrough articles that go against existing theory are accepted by the scientific community faster if they provide an answer to a long-standing question, for example. This is likely, as the question-driven approach of such articles may provide more legitimacy to the anomalous finding than if it were driven by new evidence. We therefore encourage others to extend our analysis to a larger set of breakthrough articles, potentially also including a broader range of scientific disciplines.

ACKNOWLEDGMENTS

The authors thank Kyle Siler for help in coding articles and Iris Wanzenböck for useful comments on drafts.

AUTHOR CONTRIBUTIONS

M. Wuestman: Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing—original draft. J. Hoekman: Conceptualization, Methodology, Supervision, Validation, Writing—review & editing. K. Frenken: Funding acquisition, Conceptualization, Supervision, Methodology, Validation, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

M. Wuestman and K. Frenken are financed by the Vici grant (453-14-014), awarded by the Netherlands Organisation for Scientific Research (NWO). J. Hoekman is financed by a Veni grant (451-15-037), also awarded by NWO.

DATA AVAILABILITY

The main data set for the article is available at <https://doi.org/10.6084/m9.figshare.12530042>. The citation data cannot be made publicly available due to the licensing contract terms of WoS.

REFERENCES

- AAAS. (2018). 2018 Breakthrough of the Year.
- Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159–170.
- Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.
- Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., ... Zollikofer, C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*, 418(6899), 801. <https://doi.org/10.1038/nature01005>
- Cameron, C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.
- Cavalli-Sforza, L., & Feldman, M. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton: Princeton University Press.
- Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, 20(1). <https://doi.org/10.1177/001316446002000104>
- Copeland, S. (2019). On serendipity in science: Discovery at the intersection of chance and wisdom. *Synthese*, 196(6), 2385–2406. <https://doi.org/10.1007/s11229-017-1544-3>
- Evans, J. P. (2016). (Mis)understanding science: The problem with scientific breakthroughs. *Hastings Center Report*, 46(5), 11–13. <https://doi.org/10.1002/hast.611>
- Gabunia, L., Vekua, A., Lordkipanidze, D., Swisher, C. C., Ferring, R., Justus, A., ... Mouskhelishvili, A. (2000). Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: Taxonomy, geological setting, and age. *Science*, 288(5468), 1019–1025. <https://doi.org/10.1126/science.288.5468.1019>
- Geijsen, N., Horoschak, M., Kim, K., Gribnau, J., Egan, K., & Daley, G. Q. (2004). Derivation of embryonic germ cells and male gametes from embryonic stem cells. *Nature*, 427(6970), 148–154. <https://doi.org/10.1038/nature02247>
- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979), 710–722. <https://doi.org/10.1126/science.1188021>
- Grumet, G. W. (2008). Insubordination and genius: Galileo, Darwin, Pasteur, Einstein, and Pauling. *Psychological Reports*, 102(3), 819–847. <https://doi.org/10.2466/PRO.102.3.819-847>
- Hage, J., & Mote, J. (2010). Transformational organizations and a burst of scientific breakthroughs: The Institut Pasteur and biomedicine, 1889–1919. *Social Science History*, 34(1), 13–46. <https://doi.org/10.1017/S0145553200014061>
- Helmer, R. L., & SNO Collaboration. (2002). First results from the Sudbury Neutrino Observatory. *Nuclear Physics B – Proceedings Supplements*, 111(1–3), 122–127.
- Hilgard, J., & Jamieson, K. H. (2017). Does a scientific breakthrough increase confidence in science? News of a Zika vaccine and trust in science. *Science Communication*, 39(4), 548–560. <https://doi.org/10.1177/1075547017719075>
- Hinrichs, M. M., Seager, T. P., Tracy, S. J., & Hannah, M. A. (2017). Innovation in the Knowledge Age: Implications for collaborative

- science. *Environment Systems and Decisions*, 37(2), 144–155. <https://doi.org/10.1007/s10669-016-9610-9>
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- Koshland, D. E. (2007). The Cha-Cha-Cha theory of scientific discovery. *Science*, 317(5839), 761–762. <https://doi.org/10.1126/science.1147166>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Structure (Vol. 2). <https://doi.org/10.1046/j.1440-1614.2002.t01-5-01102a.x>
- Leonhardt, U. (2006). Optical conformal mapping. *Science*, 312(June), 1777–1781. <https://doi.org/10.1126/science.1126493>
- Logan, D. C. (2009). Known knows, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, 60(3), 712–714.
- Marx, W., & Bornmann, L. (2013). The emergence of plate tectonics and the Kuhnian model of paradigm shift: A bibliometric case study based on the Anna Karenina principle. *Scientometrics*, 94(2), 595–614. <https://doi.org/10.1007/s11192-012-0741-6>
- Meyers, M. A. (2011). *Happy accidents: Serendipity in major medical breakthroughs in the twentieth century*. New York: Arcade Publishing.
- Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hot-spot. *Science Advances*, 3(4), e1601315. <https://doi.org/10.1126/sciadv.1601315>
- Myers, G. (1985). Texts as knowledge claims: the social construction of two biology articles. *Social Studies of Science*, 15(4), 593–630.
- Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change* (Vol. 93). Cambridge, MA: Belknap. <https://doi.org/10.2307/2232409>
- Palmer, D. M., Barthelmy, S., Gehrels, N., Kippen, R. M., & Cayton, T. (2005). A giant gamma-ray flare from the magnetar SGR 1806-20. *Nature*, 434, 1107–1109. <https://doi.org/10.1038/nature03525>
- Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81(1), 49–55. <https://doi.org/10.1016/j.techfore.2012.09.017>
- Saldaña, J. (2015). *The coding manual for qualitative researchers*. Thousand Oaks, CA: SAGE Publications Ltd.
- Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy*, 40(10), 1321–1331. <https://doi.org/10.1016/j.respol.2011.06.009>
- Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, 11(1), 46–62. <https://doi.org/10.1016/j.joi.2016.11.001>
- Springer Nature. (n.d.). Latest research and news by subject. Retrieved 30 January 2020, from <https://www.nature.com/subjects>
- Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.
- Toulmin, S. E. (1967). The evolutionary development of natural science. *American Scientist*, 55, 456–471.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99–114.
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472. <https://doi.org/10.1126/science.1240474>
- Ventegodt, S., & Merrick, J. (2004). Philosophy of science: How to identify the potential research for the day after tomorrow? *The Scientific World Journal*, 4, 483–489. <https://doi.org/10.1100/tsw.2004.103>
- Winnink, J. J., & Tijssen, R. J. W. (2014). R&D dynamics and scientific breakthroughs in HIV/AIDS drugs development: The case of integrase inhibitors. *Scientometrics*, 101(1), 1–16. <https://doi.org/10.1007/s11192-014-1330-7>
- Wu, L., Wang, D., & Evans, J. A. (2019). Large teams have developed science and technology; small teams have disrupted it. *Nature*, 566, 378–382. <https://doi.org/10.2139/ssrn.3034125>
- Yaqub, O. (2018). Serendipity: Towards a taxonomy and a theory. *Research Policy*, 47(1), 169–179. <https://doi.org/10.1016/j.respol.2017.10.007>
- Zeng, C. J., Qi, E. P., Li, S. L., Stanley, H. E., & Ye, F. Y. (2017). Statistical characteristics of breakthrough discoveries in science using the metaphor of black and white swans. *Physica A: Statistical Mechanics and Its Applications*, 487, 40–46. <https://doi.org/10.1016/j.physa.2017.05.041>

APPENDIX

Table A1. Examples and key phrases per dimension state

Dimension	State	Example	Key phrases	
A: The discovery is driven by a question, or by a research object	A1: Driven by question	2006 Cloaking Technology 10.1126/science.1126493	"This study develops a general recipe for the design of media that create perfect invisibility within the accuracy of geometrical optics"	"It has been a long-standing question whether [...]"; "[...] remains largely unknown"; "We hypothesize that [...]"; "We do not understand the working of [...]"
	A0: Driven by research object	2008 Seeing Exoplanets 10.1126/science.1166585	"High-contrast observations with the Keck and Gemini telescopes have revealed three planets orbiting the star HR 8799 [...]"	"We report the discovery of [...]"; "The discovery of [...] sheds new light upon [...]"
B: The discovery introduces a new question/research object, or contributes to a known question/research object	B1: New	New question 2009 Live Long and Prosper 10.1038/nature08221	"Inhibition of the TOR signalling pathway by genetic or pharmacological intervention extends lifespan in invertebrates, [...]. However, whether inhibition of mTOR signalling can extend life in a mammalian species was unknown."	"We raise the question whether [...]"; "Since we know [...] and [...], it follows logically to ask [...]", "we hypothesize that [...]"
		New research object 2002 The Tournai Fossil 10.1038/nature.00879	"Here we report the discovery of six hominid specimens from Chad, central Africa."	"We report the discovery of [...]"; "On [date] we have observed..."
	B0: Known	Known question 2001 Carbon Consensus 10.1126/science.1057320	"Despite widespread consensus about the existence of a terrestrial carbon sink [...], the size, spatial distribution, and cause of the sink remain uncertain (refs)."	"Earlier research has raised the question whether [...]"; "It has been a long-standing question whether [...]"
		Known research object 2006 Tiktaalik Fossil Fish 10.1038/nature04637	"Here we describe the pectoral appendage of a member of the sister group of tetrapods, Tiktaalik roseae (reported elsewhere, red.), which is morphologically and functionally transitional between a fin and a limb."	"The discovery of [...], by [ref], provides us with new insights into [...]"

Downloaded from http://direct.mit.edu/qs/article-pdf/1/3/1203/1869971/qs_a_00079.pdf by guest on 21 September 2021

Table A1. (continued)

Dimension	State		Example	Key phrases
C: The question/ research object is against, in line with state-of-the-art literature	C1: Against literature	2000 New Cells for Old 10.1126/ science.288. 5471.1660	“The differentiation potential of stem cells in tissues of the adult has been thought to be limited to cell lineages present in the organ from which they were derived. [...] We show here that neural stem cells from the adult mouse brain can [...] give rise to cells of all germ layers.”	“[...] which is against the theory of [...]”, “There are two competing theories, and we present evidence against one in support of the other”
	C0: In line with literature	2006 Shrinking Ice 10.1029/ 2006GL026369	“We estimate mass trends over Antarctica using gravity variations [...], similar to a recent estimate of ice mass loss from satellite altimetry and remote sensing data.”	“[...] which confirms the theory of [...]”; “[...] so that we can extend the model introduced by [...]”

Table A2. Examples and key phrases for reported considerations of use

Considerations of use		Examples	Key phrases
Applied	2009 Live Long and Prosper 10.1038/ nature08221	“These findings have implications for further development of interventions targeting mTOR for the treatment and prevention of age-related diseases.”	“These findings will help us develop [...]”; “Our findings can be used to design [...]”, “Our findings can provide a platform to cure [...]”, “Potential uses of our findings are [...]”
Fundamental	2006 Biodiversity and Speciation 10.1126/ science.1126121	“This work has specific implications for understanding the evolutionary mechanisms responsible for adaptive phenotypic change.”	“These findings will help us understand [...]”; “These findings provide insight into [...]”

Table A3. Coefficients for negative binomial regression models, with noncumulative citation counts

	cit_1y (9)	cit_2y (10)	cit_5y (11)	cit_10y (12)	cit_1y (13)	cit_2y (14)	cit_5y (15)	cit_10y (16)
# authors (log)	0.343*** (0.043)	0.314*** (0.045)	0.202*** (0.061)	0.082 (0.100)	0.324*** (0.044)	0.293*** (0.047)	0.182*** (0.064)	0.065 (0.105)
Dimension A	0.570*** (0.157)	0.746*** (0.162)	1.031*** (0.222)	0.715** (0.299)	0.399** (0.162)	0.662*** (0.172)	0.978*** (0.239)	0.692** (0.316)
Dimension B	0.135 (0.109)	0.099 (0.112)	0.090 (0.154)	-0.027 (0.212)	0.077 (0.110)	0.069 (0.117)	0.053 (0.162)	-0.106 (0.225)
Dimension C	0.0004 (0.166)	0.126 (0.172)	0.053 (0.234)	0.241 (0.314)	-0.046 (0.164)	0.114 (0.173)	0.041 (0.240)	0.259 (0.319)
Constant	2.915*** (0.197)	3.062*** (0.204)	3.076*** (0.279)	3.404*** (0.388)	3.276*** (0.223)	3.270*** (0.237)	3.233*** (0.329)	3.602*** (0.457)
Discipline	excluded	excluded	excluded	excluded	included	included	included	included
Observations	335	335	335	244	335	335	335	244
Log Likelihood	-1,760	-1,837	-1,829	-1,265	-1,681	-1,757	-1,752	-1,192
Akaike Inf. Crit.	3,531	3,685	3,668	2,539	3,377	3,531	3,519	2,400

Note: *** $p < .001$, ** $p < .01$, * $p < .05$.

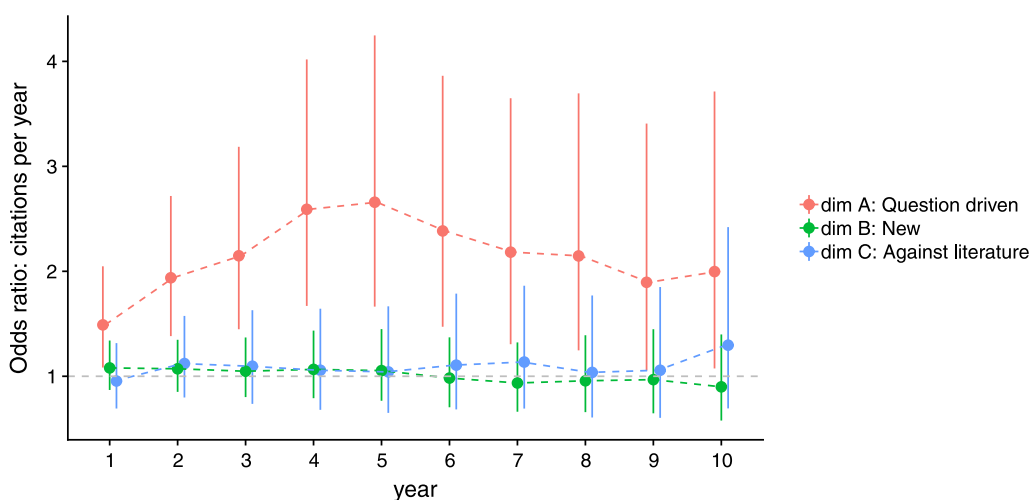


Figure A1. Incidence rate for noncumulative citations after N years, after controlling for the log of the number of authors and discipline.