Searching for new breakthroughs in science: how effective are computerised detection algorithms?☆

J.J. Winnink[a,*], Robert J.W. Tijssen[a,b], A.F.J. van Raan[a]

[a] *Centre for Science and Technology Studies (CWTS), Leiden University, Kolffpad 1, P.O. Box 905, 2300 AX Leiden, The Netherlands*
[b] *DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University, Private Bag X1, Matieland 7602, South Africa*

Abstract

In this study we design, develop, implement and test an analytical framework and measurement model to detect scientific discoveries with 'breakthrough' characteristics. To do so we have developed a series of computerized search algorithms that data mine large quantities of research publications. These algorithms facilitate early-stage detection of 'breakout' papers that emerge as highly cited and distinctive and are considered to be potential breakthroughs. Combining computer-aided data mining with decision heuristics, enabled us to assess structural changes within citation patterns with the international scientific literature. In our case studies we applied a citation impact time window of 24–36 months after publication of each research paper.

In this paper, we report on our test results, in which five algorithms were applied to the entire Web of Science database. We analysed the citation impact patterns of all research articles from the period 1990–1994. We succeeded in detecting many papers with distinctive impact profiles (breakouts). A small subset of these breakouts is classified as 'breakthroughs': Nobel Prize research papers; papers occurring in Nature's Top-100 Most Cited Papers Ever; papers still (highly) cited by review papers or patents; or those frequently mentioned in today's social media. We also compare the outcomes of our algorithms with the results of a 'baseline' detection algorithm developed by Redner in 2005, which selects the world's most highly cited 'hot papers'.

The detection rates of the algorithms vary, but overall, they present a powerful tool for tracing breakout papers in science. The wider applicability of these algorithms, across all science fields, has not yet been ascertained. Whether or not our early-stage breakout papers present a 'breakthrough' remains a matter of opinion, where input from subject experts is needed for verification and confirmation, but our detection approach certain helps to limit the search domain to trace and track important emerging topics in science.

Keywords: scientific breakthroughs, computerized search algorithms, early stage detection, citation impact patterns, Nobel Prizes

1. Introduction

Scientific and scholarly research may result in a new discovery[1]. The nature and impact of such a discovery on the cognitive structure and evolution of science may vary considerably. Some of those discoveries, each showing a major impact on future scientific research, are considered to signal possible breaches, focus shifts, or even turning points in science. The term *breakthrough* is usually used for those discoveries that have such a major impact on science. The impact of discoveries may extend beyond the domain of science and may be crucial steps towards technological applications, and to innovations and products. In line with Grupp and Schmoch (1992) several other well-known studies for instance the *Hindsight* study (Isenson, 1969), the studies conducted by Jewkes et al. (1958), studies (Heilbron, 1972; IIT Research Institute, 1968, 1969) by the Illinois Institute of Technology Research investigating the research and development process leading to innovation, the *Battelle* study (Globe et al., 1973), the *Retrosight* project (Wooding, 2007) and also the TRACES study (Walsh, 1973) searched for the impact of scientific discoveries on the development of technology[2]. A conclusion in all these and other studies is that it can take many years before a scientific discovery finds its way into new or adapted technology[3]. Scientific discoveries and their incorporation in technology are often interlinked in complex ways within research and development (R&D)[4] systems, and may span several years, decades, or even centuries.

Given the vast number of scholarly publications published each year an automated computerised selection system might be a preferable method to harvest databases with bibliographic data of scholarly publications and to search for high-impact publications. Such a generalized and transparent method should facilitate the early and unbiased detection of potentially important new directions in science and technology. An objective method, consisting of one or more algorithms, is relevant as human beings who carry out the evaluation of new developments might be forced to follow a set of strict protocols. The role of these protocols is to prevent preconceptions that could influence this process of evaluation. Fore-

☆These authors contributed equally to this work.
*Corresponding author
Email address: winninkjj@cwts.leidenuniv.nl (J.J. Winnink)

[1]Discovery - An observation or finding of something unknown prior to that discovery

[2]Technology – the application of scientific knowledge for practical purposes

[3]An example is Graphene. Based on theoretical physical calculations the properties to be expected for a material currently known as 'graphene' were presented in 1947 by Wallace (1947). It was not until 2004 however, with the publication by Novoselov et al. (2004) when 'freestanding' graphene became a reality and the predicted properties could be experimentally verified. The Nobel Prize Physics was awarded in 2010 to Konstantin Novoselov and André Geim for this discovery

[4]R&D — general term for activities in connection with corporate or governmental innovation

casting the changes that discoveries may bring about, and monitoring or nowcasting[5] the evolution of emerging areas in science or technology, presents us with a series of conceptual and methodological challenges.

In this paper we focus on methods to detect discoveries that change the fabric of science itself, more specifically the immediate impacts within the first two to three years after the discovery was published. Such early-stage detection of major discoveries is relevant not only to scientists themselves, but also to government policy-makers and corporate R&D executives as they may signal significant focus shifts in industrial R&D systems. On the basis of such information funding strategies can be adapted knowing a possible major new development exists. Policy-makers and funding agencies have a particular interest in knowing in which direction research[6] and innovation[7] are heading to gain or sustain economic growth and prosperity, or to allocate scarce resources for R&D. These R&D decision-makers usually only oversee the areas of science and technology they focus on, and therefore they might easily miss or misinterpret relevant (fast moving) developments outside their focal area. The newest developments with possible large and immediate impacts on 'upstream' science and technology in later 'downstream' stages of development are of particular importance.

This paper further discusses on the identification at early stage publications that have the potential to stimulate areas to evolve into 'hot spots' in science. In this paper the research objectives, the theoretical framework that is used as a basis, the methodology and data sources, empirical results, conclusions and insights are presented. Detailed additional supporting information in relation to this research can be obtained from the authors and can also be found in Winnink (2017).

2. Method and data sources

2.1. Theoretical and conceptual framework

Science as a dynamic system. Science can be considered a dynamic system[8] in which scholars and their research activities play a dominant role, and discoveries can act as events that change the nature, shape or direction of scientific progress — either in terms of new knowledge production, or an interpretation or reinterpretation of existing knowledge, ideas and know-how. In general, systems operate in the vicinity of a certain equilibrium state and are considered to be stable unless factors force the system to undergo larger-than-usual

[5] Nowcasting - the activity of estimating the current situation on the basis of historic data

[6] Research – studious inquiry or examination; especially: investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws. Source: http://www.merriam-webster.com/dictionary/research

[7] Innovation – the act or process of introducing new ideas, devices, or methods. Source: http://www.merriam-webster.com/dictionary/innovation

[8] System – a group of related parts that move or work together Source: http://www.merriam-webster.com/dictionary/system

changes and to enter a state from which it cannot readily return to the previous situation. Such changes from one stable state to another are also called 'phase transition' or 'phase change', in analogy to comparable processes in physics, chemistry and biology. The dynamic behaviour of complex systems is covered extensively in the scholarly literature for instance (von Bertalanffy, 1969). All systems, not just the large ones, can undergo irreversible[9] changes (Mandelbrot, 1982). Several empirical studies (Scheffer et al., 2009; Scheffer, 2010, 2009; Lade and Gross, 2012) have shown that dynamic systems can transmit early-warning signals indicating a 'phase transition' is about to happen; a transition to a new state in which it stays until a new event forces the system to move to yet another state. Such a major transition stands out between the more common 'minor' changes a system undergoes frequently. Whether or not a scientific discovery should be considered a minor or major change — a breakthrough or not — has been a topic of study and academic debate during the last 50 years.

Progress in science. There is a general notion that science progresses on the basis of work done by scholars, researchers and scientists that builds on prior achievements (often by others); as described by the motto "If I have seen further it is by standing on the shoulders of giants"[10]. The evolution of science, however, does not follow a linear, continuous, cumulative unified path, which is the impression of the development of science as it emerges from textbooks (Kuhn, 1962), where the knowledge is ordered in such a way that it can serve education. Kuhn distinguishes 'normal' science and 'revolutionary' science and argues that the development of science alternates between these two states. In normal science discoveries fit within an existing paradigm[11] and are expected[12]. Revolutionary science deals with those discoveries that are at odds with the then existing paradigm.

Normal science, in Kuhnian terminology, is scientific research conducted within a single paradigm. Within normal science the foundations of the paradigms and the paradigms themselves are not argued, and science research functions as a 'puzzle-solving' activity inside a framework of common understandings and starting points. At the point when *tension* between the then current paradigm and observations from scientific research occurs, a new paradigm might come into existence, in which case a 'paradigm shift' can be observed. Wray (2011, p.202) argues: "...According to Kuhn's mature view, a new theory is developed in a field in an effort to account for an anomaly that the accepted theory was unfit to account

[9] Without external influence the system is incapable of returning to the previous condition or state

[10] This metaphor is usually attributed to Sir Isaac Newton, but should be ascribed to Bernard of Chartres as it was first recorded in the 12th century (Merton, 1965, p.267)

[11] Kuhn (1962) defines a paradigm as "...that which the members of a scientific community, and they only share..."

[12] Although there is a sense that a discovery is forthcoming the exact moment it will happen is uncertain

for ...". The new paradigm enables the resolution of previously unsolvable problems and replaces the old one. Paradigm shifts proliferate slowly as the relevant scientific community needs to be convinced to alter its views and approaches. This process will go on forever. Kuhn's observation of discontinuities in the development of science is now widely accepted. Results of scientific research that can only be explained by changing an existing paradigm are characteristic for revolutionary science, according to Kuhn. Radical novel approaches, new information and discoveries, which are incompatible with the current dominant theoretical framework and beliefs within a science field, may suddenly appear on the scene and revolutionize the cognitive structure of that field (Andersen et al., 2006). These are the 'phase transitions' that have a large impact on science within a relatively brief span of time.

Discoveries in science. Identical or related discoveries frequently come in a manifold — "...It is an interesting phenomenon that many inventions[13] have been made two or more times by different inventors, each working without knowledge of the other's research..." (Ogburn and Thomas, 1922). Such 'multiple discoveries' may differ in appearance, and occur at different points in time, or at different geographical locations. Merton (1961, Ch.II, p.478), who confirms the observations made by Ogburn and Thomas (1922) expands on the notion of manifold discoveries concluding that "...singletons, rather than multiples, are the exception requiring distinctive explanation and that discoveries in science are, in principle, potential multiples...". Merton (1961, p.480) also refers to his study on historical incidents of multiple discoveries in which he reported on the occurrence of up to five and six-fold discoveries. Price (1963, p.65–66) also discusses this phenomenon and links it with Kuhn's concept of normal science in which discoveries in a sense are to be 'expected' from time to time. Simonton (1978, 1979) and Brannigan and Wanner (1983) analysed historic data on sequences of discoveries in science to uncover the mechanism behind the phenomenon of multiple discoveries. Brannigan and Wanner (1983) conclude that of the several possible stochastic models that can be used to describe the distribution of the grade of multiples, models based on a Poisson distribution gives adequate results.

Scholarly communication. Scientists generally use the results achieved by other researchers as a starting point for their scientific insights, as is expressed by the already mentioned metaphor *"If I have seen farther, it is by standing on the shoulders of giants"*. A discovery can only contribute to advances in science when it is codified in a way that allows communication to others. The principal means of scholarly communication is by text; in modern-day times

[13] Nowadays the terms 'discovery' and 'invention' have distinctive and separate meanings. In the past these terms were used interchangeably, and 'invention' was also used in situations where currently the term 'discovery' is preferred

usually by research articles in scholarly or technical journals ('research publications') or in books. Price (1963, Ch.3, p.68) discusses the role of scientific publications, and concludes that "...The scientific paper therefore seems to arise out of the claim staking brought on by so much overlapping endeavour. The social origin is the desire of each man to record his claim and to reserve if for himself..." Sharing and claiming research findings is a major reason for communication within the scientific community. Communication in science takes place in several forms. Formal means of communication are scholarly publications, conference proceedings, and books. Less formal[14] ways of communication play a role within teams of collaborating researchers who have close working relations in which information sharing is obligatory for the team to be able to function optimally. Crane (1972) concludes, that as scientists rely on research results of other researchers, they group together in 'invisible colleges'. The citing-cited relations between scholarly publications form the fabric of these invisible colleges. Formal means of communication within the virtual colleges are publications (Lievrouw, 1989). Price (1965) argues that the pattern of bibliographic references reflects the nature of the scientific research front. Lievrouw (1989, p.616) examines the relationship of bibliometric techniques, especially citation analysis, with communication theory and research, and argues "...However, it[15] is of particular interest here because it is possibly the best-known model of scientific communication...".

Citation analysis — e.g. Moed et al. (2004) — acts as an important framework for the analysis of various aspects of the scientific community, and is a central theme in bibliometrics[16] Developments in science can be monitored using citation relations between scholarly publications. Citation relations between patent publications and scholarly publications provide an — albeit partial — view on the influence of scientific research on technological evolution. Citations are biased in the sense that they are influenced by several mechanisms that are not directly related to the contents of the publication. Price (1963, p.87) mentions that in certain situations where the results of team research are reported "...The participating physicists are not mentioned, not even in a footnote...". Merton (1968) points to psychosocial conditions that have an impact on citation behaviour, for instance already eminent researchers are given disproportionate credit in some cases. Crane (1972, p.83) concludes that social factors within a research field have an effect on the diffusion of knowledge in the field, and that these factors furthermore determine which information is to be used in later publications. Notwithstanding the limitations, citation relations can be used as a proxy to reveal the

[14] Data from social media is not taken into account in this study as this is (1) a recent development and (2) the value for the analysis such as those carried out in this study is not yet evident

[15] [Added by the author] with 'it' Liefrouw refers to the concept of 'invisible colleges'

[16] Pritchard (1969, p.349) defines 'bibliometrics' as "...the application of mathematics and statistical methods to books and other media of communication...", and by Broadus (1987, p.376) as "... the quantitative study of physical published units, or of bibliographic units, or of the surrogates for either..."

evolution of science and technology.

The bibliographic information of scholarly publications is used in this study as a major information source. These publications do not form a homogeneous group as they present various forms of dissemination of information between scholars. A document type is assigned to each scholarly publication when the accompanying bibliographic information is stored in a database. One of the assigned classes is 'article'. Articles are considered to be the publications that contain the results of original scientific research. These publications are often multi page publications published in a scientific journal. The concept of an article seems obvious at first sight but consists of several types of publications and contains multi page publications as well as shorter publications known as, for instance, 'letters to editor' or 'opinion letters'. Such shorter publications may also contain the results of original research or can contain original ideas[17].

Discoveries and breakthroughs. Scientific practice can be seen as a system that continuously undergoes changes as a result of discoveries that influence the science system. Every 'open' scientific discovery - one that is properly documented and communicated to others within the relevant research community - is likely to have an impact, although (at first) possibly negligible, on (r)evolutionary changes in science. The discoverers and other subject experts[18] are able, usually with the benefit of hindsight, to identify and value those impacts after a period of time.

When studying the evolution of science fields, the flow of publications is often used as an approximation of the dissemination of the knowledge related to a scientific finding and of the way other researchers follow-up on this finding. Knowledge diffusion does not guarantee a continuous gradual evolution of science because the diffusion process changes parameters in the system and can therefore result in unexpected high-impact changes. Some discoveries might not be noticed for some time, for many reasons.[19] Some might even be totally neglected where the result is not properly documented or communicated – the result is forgotten or its implications overlooked. Only a small number of scientific discoveries lead to large, structural changes in science fields, and pave the way for novel insights and further productive research. For a discovery to be qualified as a distinctive 'major' discovery not only requires the judgement by a wider range of subject experts, but also needs sufficient length of time to allow extensive validation and general appreciation. Becattini et al. (2014)[20]

[17] Koshland (2007) is an example of a two-page manuscript typified as 'opinion letter' containing relevant original ideas to which the document class 'article' is assigned

[18] Subject expert - An expert in the field; someone who has specific knowledge concerning a subject

[19] The fact that a discovery might remain unnoticed by the scientific community for some time, but is later considered a breakthrough, is for instance described by Ciechanover (2009, p.2)

[20] Becattini et al. (2014) "... After 1985 about 15% of physics, 18% of chemistry, and 9% of medicine prizes were

analysed the time lag between discoveries and the awarding of a Nobel Prize and conclude that Nobel prizes are awarded only very rarely within 10 years of a discovery.

The term 'breakthrough' is usually applied to such discoveries, a term frequently used for events that are considered major discoveries. Major journals like *National Geographic, Nature* and *Science* regularly publish overviews of what they regard as the major scientific discoveries in a previous period or specific year; these lists are usually based on expert opinions. What exactly is meant by a breakthrough or major discovery is not specified, and for good reason: there is no generally accepted, let alone a universal, definition that can count on full support throughout the scientific community. In particular the notion "...new way of thinking about a problem..." is an essential property of a breakthrough put forward by Hollingsworth (2008, p.317). The term breakthrough is not only used for the nature of the transition ('What exactly did change?') but is also used for the point in time the event occurred ('When did the change occur?') or to the impact the discovery had on other systems. The fact that the same term is used for closely related phenomena is elucidated in (Hofstadter and Sander, 2013, Ch.1).

This absence of a generally accepted definition is illustrated by the fact that various synonyms are in use for the term breakthrough such as 'advance', 'development', 'step forward', 'quantum leap', 'evolution', and others. The lack of a single definition for a breakthrough complicates the identification of these phenomena. Hollingsworth (2008, p.317) defines a breakthrough as "...A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem ... This new way of thinking is highly useful to numerous scientists in addressing problems in diverse fields of science...". Hollingsworth argues further that science evolves not just through the occasional breakthroughs but also by means of numerous successive small, incremental advances. The co-existence and interplay between 'incremental' and 'breakthrough' advances is in line with Kuhn's idea that after a 'paradigm shift' (i.e. revolutionary science) has occurred 'normal science' will take over — at least for some period of time (Kuhn, 1962).

In spite of the lack of proper operationalization and identification, there is general agreement only on the fact that breakthroughs are, by definition, rare events. The precise moment such a major change occurred is even in retrospect hard to pinpoint and foretelling when such an event is likely to occur is near possible. Nonetheless, some progress is currently being made on theoretical and empirical models that may enable forecasting or prediction methods. For instance, Ball (2004) discusses the fact that systems need a certain 'critical

awarded within 10 years of the corresponding discoveries. By contrast, before 1940 about 61% of physics, 48% of chemistry, and 45% of medicine prizes were awarded within 10 years of the corresponding discoveries..."

mass' to undergo a major change. Complex dynamic systems can have 'tipping points' (Scheffer et al., 2009; Scheffer, 2010). The prediction of such tipping points before they are reached is, however, extremely difficult. Scheffer et al. (2009) concludes "...work in different scientific fields is now suggesting the existence of generic early-warning signals that may indicate for a wide class of systems whether a critical threshold is approaching...".

Breakthrough discoveries are events that have a major impact on future scientific research and can be considered a tipping point in science. Several scholars constructed theoretical models that focus on the diffusion of knowledge within the scientific community and connect this knowledge diffusion with the occurrence of discoveries. Andersen et al. (2006) focus on the cognitive changes that occur in science when a paradigm shift occurs. In the publications of Bettencourt and colleagues (Bettencourt et al., 2009; Bettencourt and Kaiser, 2011, 2015) a *percolation model* describing the development of science is proposed. Bonaccorsi (2008, 2010) hypothesises that new science fields that came into existence after the 1970s follow a different evolutionary development path compared to already established sciences. Chen et al. (2009) propose an explanatory and computational theory of transformative discoveries in science. Cintron-Arias et al. (2005) tested mean-field deterministic epidemic models to describe knowledge diffusion. A catastrophe model to develop a formal non-linear model of scientific change in concordance with Kuhn's hypotheses is put forward by Perla and Carifio (2005). Sung (2008) shows that experiments play a crucial role in formulating an explanation of the RNAi anomaly. Vitanov and Ausloos (2012) focus on the uses of compartmental epidemic models[21] — Lottka-Volterra's model and others — to describe technology diffusion. These and other theoretical models contain conceptual descriptions of the evolution of the science system and can therefore be used to identify areas that evolve into hot spots.

Characterizing discoveries. Time not only picks the winners, it also unmasks discoveries that turned out to be hypes[22], hoaxes and frauds. An example of a hoax is the claim for the existence of nuclear fusion at room temperature — 'cold fusion' (Fleischmann and Pons, 1989). This claim was almost immediately criticized, and it was concluded (Dmitriyeva et al., 2012) that "...According to our calculations, the experimentally measured excess heat can be accounted for fully by this chemical reaction...". The Korean researcher Hwang Woo-Suk was considered one of the pioneering experts in the field of stem cell research until a publication by Cyranoski (2004) uncovered Hwang's fraudulent research. The increasing incidence of retraction of scientific publications (Cokol et al., 2008) blurs the picture of the

[21] Compartmental epidemiological models with a name based on the specific compartment structure of the model, e.g. SIS, SIR, SEIR

[22] Hype – extravagant or intensive publicity or promotion (source: Oxford Dictionary of English, 2nd edition)

Table 1: Cha-Cha-Cha typology of discoveries

Koshland type	Type of discovery	Kuhnian type	Characterisation
Charge	These discoveries solve problems that are quite obvious, but in which the way to solve the problem is not so clear	*Normal* science	'…In these discoveries, the scientist is called on, as Nobel laureate Albert Szent-Györgyi put is "to see what everyone else has seen and think what no one else has thought before …"'(Koshland, 2007, p.761)
Challenge	These discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time.	*Revolutionary* science	"Sometimes the discoverer sees the anomalies and also provides the solution. Sometimes many people perceive the anomalies, but they wait for the discoverer to provide a new concept."(Koshland, 2007, p.761)
Chance	These discoveries are often called serendipitous	*Revolutionary* science	'finding the unsought' (van Andel, 1994), like the discoveries of penicillin or Teflon® (Koshland, 2007)

bibliographic data. Clearly the term 'breakthrough discovery' should be used with caution. So, perhaps it is not surprising that in the academic literature there is a lack of convincing typologies or helpful classification systems of scientific discoveries.

One of the exceptions is the 'Cha-Cha-Cha' theory developed by Koshland (2007), who classifies scientific discoveries into three distinct classes based on the nature of the discovery in relation to already existing scientific knowledge: ***Cha**rge, **Cha**llenge* and ***Cha**nce*. Koshland's (2007) classification (see Table 1 on page 10), focusing on how a discovery is different from the then existing scientific knowledge, is just one way of classifying discoveries. Scientific discoveries can be classified on the basis of various characteristics. Redner (2005) for instance classifies discoveries that are documented and presented in a scholarly publication, according to citations of those publications by other scholars. We will return to Redner's interesting approach, which classifies discoveries as either *non-breakthrough* or a *breakthrough*. Discoveries of type Charge are the most common.

Collaboration and research teams. Discoveries are not only about those by individual, prize-winning 'giant' researchers and scholars, those who allegedly "stand on the shoulders" of other preceding giants, they are also about joint efforts and research collaboration between individuals benefiting from each other's knowledge, inspiration and know-how. Watson & Crick's discovery of DNA in the 1950s is probably the most well known example in contemporary science. As modern science itself has become much more collaborative (Wuchty et al., 2007), especially 'big science' based on large shared research facilities (Price, 1963), the same is likely to apply to scientific breakthroughs. In an empirical study Uzzi et al. (2013) conclude that teams are 37.7% more likely than solo authors to insert novel combinations of prior work into familiar knowledge domains. Publications with such novel and unusual combinations are rare but are twice as likely to become highly cited works.

Analysing team collaboration in general one of the main conclusions drawn by Uzzi and Spiro (2005, p.492) is "...Small world networks[23] do benefit performance but only up to a threshold, after which the positive effects of small worlds reverse...". Whitfield (2008, p.720–723) concludes that research is becoming more and more a matter of team activity and the contribution of single authors to science is dwindling. Green and Brendsel (2008) respond that this observation might be correct, but "...Lightning can still strike the solitary explorer whose mind is prepared...". These results are in line with Wuchty et al. (2007) who conclude that knowledge-producing teams increasingly dominate over solo authors, and that their publications are more frequently cited; this citation advantage increases over time. These authors furthermore conclude that the process of knowledge creation has undergone a fundamental change in moving from research conducted by individual researchers to research carried out by teams of researchers.

Burt (2004) concludes that the opinion and behaviour of people belonging to a group are more homogeneous within a group than between groups. Individuals who are part of multiple groups can therefore bridge cognitive gaps between groups and come up with solutions that otherwise might be unseen. Guimerà et al. (2005) conclude that the forming of a large group of practitioners can be described as a 'phase transition' after having analysed more than 4 million publications issued in a period of more than 30 years. Jones et al. (2008, p.1261) conclude that collaboration in science is increasingly becoming composed of co-operations spanning university boundaries. According to Skilton (2009) articles co-authored by teams that include frequently cited scholars and teams whose members have diverse disciplinary backgrounds are cited more often. Weinberg (1970, p.1056) argues that the formation of large interdisciplinary teams centred on pieces of expensive equipment causes the increasing importance of team science, especially since World War II. Such research teams are said to be part of 'big science' (Price, 1963). Work groups construct a common group identity over time through the process of value convergence between group members (Meeussen et al., 2014). Bettencourt et al. (2008) analyse the quantitative social structures of collaboration that develop as new scientific fields emerge. An increased interaction between scientists exploring different aspects of a problem creates new concepts, techniques and a shared research programs resulting in successful new fields.

Research teams are not fixed structures and evolve over time. Tuckman (1965) introduces what has become known as the standard model of small group development in which four stages are distinguished. McGrew et al. (1999) extend Tuckman's model with three declining phases to create a model that describes both the formation and the decay of small groups.

[23][Added by the author] A small-world network is a type of mathematical graph in which most nodes are not neighbours of one another, but most nodes can be reached from every other node by a small number of hops or steps. Source: https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)

Bettencourt and Kaiser (2011) point to the difficulty in defining and comparing science fields and observe that there seems to be a general sense that the different fields undergo similar stages of development. Of particular importance in a field's history are the moments at which conceptual and technical unification allows the widespread exchange of ideas and collaboration. These moments mark the point in time when the networks of collaboration between scholars show the analogue of a percolation phenomenon, and develop a giant connected component containing most authors of a conceptual framework.

2.2. Analytical framework

Facing the methodological challenge. Bibliographic information within a research publication that first describes a discovery may refer to it's relevance or anticipated relevance for scientific progress, but its true impact depends on its reception and implementation over time. Subsequent research publications, referring to the discovery and its list of literature references ('citations'), will reflect and reveal reactions from peers in the scientific community to the breakthrough work. As a consequence, bibliographic information can only help identify breakthroughs in those cases where these scholarly publications receive exceptional scores on citation-impact metrics[24]. Publications with a large 'citation impact', or those that are immediately cited, are more likely to be seen - with hindsight - as breakthrough publications by the scientific community. But reaching such shared opinion takes time - often many years or decades. In this thesis, we will refer to highly cited publications that have not (yet) acquired breakthrough status as *breakout* publications, or *breakthrough by proxy*[25]. Only after sufficient time has elapsed, and with the benefit of expert opinions, will some breakouts be considered a breakthrough. As put forward in Section 2.3 on page 15 this study addresses the following methodological challenge:

> "Is it possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool with a range of practical applications for early detection of breakthroughs in worldwide science?"

In a first step towards operationalization[26] we will focus our attention on identifying breakout publications that are characterised by specific citation impact profiles. To do so, this study focuses on the observable effects of discoveries on the research community, guided by the research question:

> "What kind of detectable evidence do discoveries leave behind in the research literature in terms of sudden changes and distinctive structural developments?"

[24] Metric – a technical system or standard of measurement
[25] This group is further broken down into subcategories as explained in Section 3.4
[26] Operationalization is the process of strictly defining variables into measurable factors

Bettencourt et al. (2009, p.220) hypothesize "...there is a universal character in discoveries..." and argue that circumstantial evidence for the existence of such universal characteristics is also supported by several other studies, such as (Gerstein and Douglas, 2007; Uzzi and Spiro, 2005; Leskovec et al., 2005). Chen et al. (2009) introduce an explanatory and computational theory of 'transformative discoveries' science based on the central premise of the connection of disparate areas of knowledge is introduced. Their theory explains the nature of these discoveries, and also characterizes the subsequent diffusion process. According to the authors the primary value of the theory is that it provides both a computational model of intellectual growth, and concrete and constructive explanations of where insightful inspirations for transformative scientific discoveries can be found.

Tapping into universal characteristics therefore opens up the possibility of designing early-detection models of breakouts and breakthroughs, either models based on small-scale case studies or those derived from large-scale quantitative analysis. Julius et al. (1977) is an early example of the former, using expert knowledge only. The aim of this study is to tackle this question systematically and in a large-scale 'macro-level' fashion, i.e. scanning world science for breakouts. Obviously, one cannot rely on 'micro-level' individual expert judgements (or expert panels) to identify and check each and every of the hundreds or thousands of potential breakthroughs. External observers and analysts, who are not experts in the field under study, will have to resort to other information sources — notably the citations between research publications. Of course, these citations are also expert based, albeit indirectly: each citation from a fellow scientists or scholar to that specific publication describing the discovery can be seen as a 'vote of relevance', an expert-based confirmation that the cited work has been noted or had an impact on follow-up scientific research.

The key methodological challenge is to develop citation-based early-detection algorithms that enable large-scale scanning of the global scientific literature — computerised algorithms to identify at an early stage those scientific publications that have, or are likely to have, an above average impact on science. For a computational perspective, an expanding set of citation-based algorithms all build on earlier research methodologies aimed at identifying and monitoring 'emerging topics', 'emerging technologies' or 'research fronts' in scientific progress, technological development, or R&D-based innovations. These 'tracing and tracking' methods do not focus on individual publications, but rather on large sets of published documents (usually research publications and/or patents); they also tend to adopt longer time periods. The US researcher Henry Small pioneered the large-scale analytical approach in the 1970s. He introduced the notion that rapid shifts in research focus, as identified in the scholarly research literature, could be regarded as a signal of 'revolutionary' change (Small, 1977). More than 30 years later, his research program is still on-going — Henry Small and his

colleagues combine direct citations and co-citations helps to adequately identify emerging topics (Small et al., 2013).

The citation-based algorithms introduced in this paper are closely related to work by Redner (2005) who classified discoveries based on the number of citations of the publications. More recently, Baumgartner and Leydesdorff (2014) applied 'group-based trajectory modelling' to citation curves of research publications. Ponomarev et al. (2012) focus on citation patterns in combination with statistical modelling, while a follow-up study by Ponomarev et al. (2014) focuses on the effects of interdisciplinarity in the subject categories, and geographical diversity. Schneider and Costas (2017) also use citation-based methods to detect potential breakthrough publications. Wang et al. (2016) introduce a measure for the combinatorial novelty of a paper to identify those that are likely to have an above average or even high-impact. The approach adopted in this study differs significantly as it not only takes as its leading principle the number of citations a publication receives, but also focuses on the dynamic influence a publication has on the scientific community. This dynamic influence is expressed not only in the number of citations but also in the sources of the citations, like for instance authors, science fields, and clustering of citations. Another difference is the focus on the period from the publication of a paper until three years later.

Designing the early-stage detection algorithms. The research in this study is, as mentioned earlier, rooted in the assumption that bibliographic information for scholarly publications can be used as a proxy to analyse and monitor (sudden) developments in science. Citation links between publications form the basis for measurement. These citation links form the citation profile of a publication that varies over time as a publication gets more and more citations. The response of the scientific society on a publication is reflected in the number of times a publication is cited and reflects the impact a publication has on the evolution of science, as far as it can be decided on the basis of citation patterns. A basic assumption is that researchers working in the same area are able to value a discovery in relation to already existing knowledge. The impact of a publication on science is in this way linked to the way it is cited[27]. As a consequence of this approach informal communication that might take place is not taken into account.

Contrary to earlier work and the methods briefly described above, the focus in this study is on the citation impact behaviour of individual scholarly publications relatively soon after publication. The detection method covers a range of citation-based criteria, which are identified by studying the seminal research publications of generally acknowledged breakthroughs. The underlying distinctive citation impact patterns of these 'breakthrough exem-

[27] In this study the citation information is used 'as is'

plars' are unravelled and their key characteristics are used as a 'citation profile' to design computerised algorithms for searching *Challenge* and *Charge* types of breakthroughs in the global research literature. These early detection algorithms should be able to: (1) identify research publications describing discoveries that are now regarded as breakthroughs, (2) track down 'breakout' discoveries that have not yet been recognized as such.

This approach relies on systematic large-scale searches within the worldwide scholarly literature. Assembling information from large, international bibliographic databases enables external, independent analysis to identify significant short-term[28] changes in publication and citation patterns. In this study, the bibliographic is extracted from the *Web of Science Core Collection* database (abbreviated here to WoS). Further information about this information source, and relevant measurement details, are provided in Section 2.4. The analytical procedure is divided into the following seven steps:

1. Search for characteristic citation patterns of the selected breakthrough publications;
2. Selection of distinctive patterns to construct and test the search algorithms;
3. Selection of WoS-indexed research publications in the period 1990–1994 (focussing original research findings published in 'article', and 'letter' document types);
4. Determining 'optimal' citation frequency threshold values to pre-select cited publications;
5. Construction of two datasets with WoS-indexed publications published in 1990–1994, with publications that belong the top 10% most cited within two years after publication. These sets are based on two types of research subfields: (1) WoS-related *subject categories*[29] and (2) in-house defined *document clusters*[30];
6. Application of all the developed detection algorithms to both datasets;
7. Quantitative, statistical analysis of the results.

2.3. Research questions and hypothesis

We advance the hypothesis:

"It is possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool that uses bibliographic information for early detection of potential breakthroughs in science."

How generic are the algorithms we constructed in terms of their efficacy across all fields of science? This paper presents the test results, focusing on three research questions:

[28] In this study 'short-term' refers to the period 2-3 years immediately after publication of a research paper (indexed by the WoS) in which the discovery is first introduced and/or described

[29] In the WoS 251 different subject categories describing different fields in science are defined

[30] A document classification method based on citation relations between publications is developed (Waltman and van Eck, 2012) as an alternative for the WoS subject categories

1. Can the algorithms be used as a generally applicable method to identify breakout papers, and if so under what data availability conditions?
2. What are the similarities and differences between the algorithms in terms of their ability to detect breakouts?
3. Can we determine the effectiveness of each algorithm in terms of identifying breakout papers that are generally regarded as breakouts and potential breakthroughs?

2.4. Data sources

Our bibliographic database consists of research papers extracted from CWTS' in-house off-line version of Clarivate Analytics'[31] Web of Science database (WoS). From this database, we selected all 2,715,651 scientific research publications from the period 1990–1994 that were tagged with the WoS document types 'article' or 'letter'. These documents are most likely to report on 'original research'. We opted for the time period 1990–1994 to track the effect of a discovery over an extended period of time, and to verify and validate whether selected papers are currently — in retrospect — (still) regarded as breakouts or breakthroughs. For reasons of citation impact normalisation, we adopt two publication-based delineations of scientific disciplines: (1) 'Categories', the equivalent of the subject categories used in the WoS[32], and (2) 'Clusters' derived from a citation-based clustering algorithm developed at CWTS (Waltman and van Eck, 2012); we refer to this method as the 'CWTS document clustering method'. Each of the 251 Categories comprises a set of entire WoS-indexed journals; the 865 Clusters each consist of large numbers of individual research papers. WoS subject categories and CWTS document clusters represent scientific disciplines that are in line with the definition used by other scholars (Darden and Maull, 1977); we refer to both as 'discipline' in a generic way.

To narrow down our search, we selected those papers that belong to the top 10% most highly cited during the first 24 months after publication[33] per Category ('Categories') or Cluster ('Clusters') per year. Categories contains 253,558 highly cited papers and Clusters 214,827. All computations and analyses were carried out separately on both datasets.

3. Results

3.1. Breakout detection algorithms

Our algorithms meet the following general specifications. The algorithms (1) can be directly derived from data-analytical results in our case studies; (2) are systematically applicable

[31] This company comprises of the former division of Thomson Reuters responsible for the WoS that has been sold (July 2016) to two investment firms: *Onex Corporation* and *Baring Private Equity Asia*.

[32] 'http://mjl.clarivate.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D' presents information on WoS subject categories

[33] In our opinion this narrowing down of the dataset is allowed given the skewness of the citation distribution in combination with the fact that we search for potential breakthrough publications, i.e. publications that stand out, at very early stage

across the Web of Science; (3) signal a sudden significant increase in one of the parameters of a paper's citation impact pattern; (4) are systematically applicable at the level of individual research papers; and (5) can be implemented without any special pre-processing of bibliographic data. From the case studies we developed, using these criteria, the following five 'general purpose' algorithms, each representing a specific characteristic of citation impact patterns.

*Application-oriented Research Impact (*ARI*).*

The purpose of this algorithm is to identify papers that bridge 'discovery-oriented science' and 'application-oriented science' as explained in Tijssen (2010). The algorithm emerged from the case study in which we noticed remarkable, almost instantaneously, shifts over time in the ratio of citations from discovery-science papers and applied-science papers in the field of Introns (Winnink et al., 2013). The focus is on papers having a substantial list of references and are highly cited within the first 24 months after publication. The majority of the referenced papers focus on 'discovery-oriented science', whereas the citing publications focus mainly on 'application-oriented science'. Each breakout paper should meet the following selection criteria that are based on all papers in Categories and Clusters:

- Number of cited papers ≥30, this is the lower boundary for the top decile of the number of original-research papers in the reference lists;
- Number of citing papers within 24 months ≥ 49, this is the lower boundary for the top decile of the number of citations received within the first 24 months by the most-highly-cited papers;
- Number of citing papers > number of cited papers;
- Majority of the cited papers focus on 'discovery-oriented science';
- Majority of the citing papers focus on 'application-oriented research'.

*Cross-Disciplinary Impact (*CDI*).*

Captures the diffusion of citing sources among multiple research disciplines. We expect to find breakout papers that are cited by an increasingly larger number of disciplines over time. The level or cross-disciplinary impact is defined as the number of different disciplines (either Categories or Clusters) that are assigned to each of the citing papers. Given the more homogeneous disciplinary composition of each Cluster, as compared to each Category, one would expect less interdisciplinary citation flows between Clusters. This aspect is especially noticeable during the first few years after publication and then almost disappears. This algorithm is labelled CDI_{sc} when applied to the Categories dataset and CDI_{dc} when applied to the Clusters dataset. Breakout papers meet the following lower threshold values per citation time window, and that are based on the values for the 'Hazuda paper' (Hazuda et al.,

2000) , which is central in our case study of HIV/AIDS research (Winnink and Tijssen, 2014):

- Categories: 1 year: >9 citing disciplines; 2 years >17 disciplines; 3 years >24 disciplines;
- Clusters: 1 year: >2 citing disciplines; 2 years >5 disciplines; 3 years >8 disciplines.

***R**esearchers-**I**nflow **I**mpact (*RII*).*

The focus is on the influx of new researchers citing the breakout paper. Our case study on Graphene research (Winnink and Tijssen, 2015) identified papers that attract a remarkable increase in unique citing researchers. Here we expect to identify breakout papers that have an impact on an increasingly large community of research-active scholars in the research domain. Focusing on the annual number of these unique authors, who are first-authors on citing research papers, we measure the inflow rate by comparing the increase in the number of researchers at the end of the 1st year after publication, and at the end of the 3rd year. Selected papers should show an increase of at least 52 new citing first-authors. This threshold results from the increase in new citing first-authors between the end of the 1st year after publication, and at the end of the 3rd year for the paper on the Graphene discovery (Novoselov et al., 2004) that was central in the analysis presented in (Winnink and Tijssen, 2015).

***D**iscoverers-**I**ntra-group **I**mpact (*DII*).*

In our study of Ubiquitin research (Winnink et al., 2016), we found that the breakout papers that describe the scientific breakthrough received most of their citations, within the first two years, from papers co-authored by authors from the same 'core group'. The discovery is at first predominately recognized and built upon by members of the same group. This algorithm is designed to find breakout papers where many citations are from papers with authors that share co-authorship relationships with the cited authors. The following selection criteria were applied:

- 90% of the citations are 'within-group' citations;
- Within-group papers are defined as papers of which at least 66% of the authors belong to the core group. This specific lower threshold avoids the inclusion of those papers for which only one member of a small group — 3 or 4 members — is (co) author;
- The minimum size of a core group is three, which value is chosen to guarantee that in combination with the above-mentioned 66% threshold, only papers written by at least two authors of the core group are considered;
- Citations are tracked within the first two years after publication.

***R**esearch-**N**iche **I**mpact (*RNI*).*

Also originating from the Ubiquitin case study, this algorithm searches for sets of citing and

cited papers, within Categories or Clusters, with above-average rates of citation-interconnectedness. A breakout paper creates a 'citation knot', i.e. a set of papers that cite the breakout paper but also cite at least one 'auxiliary' paper with direct citation ties to the breakout paper. This closely-knit set of citing and cited papers represents a 'research niche'. The next threshold values are determined by analysing for the period 1980–1982 the network of papers citing the two breakthrough papers from 1980 that in conjunction describe the ubiquitin discovery (Winnink et al., 2016).

- The number of citations received by the breakout paper, within this niche and within the first year, is larger or equal to three times the number of interconnected papers within a citation cluster;
- The lower threshold for the number of breakout-related papers in the 'citation knot' is 8.

ARI, CDI and RII are more likely to identify Charge discoveries (i.e. solving well-known and well-defined problems — Kuhn's *normal science*), while DII and RNI are better equipped to find Challenge discoveries (i.e. explaining strange, unexpected phenomena — Kuhn's *revolutionary science*). As for Chance discoveries and breakthroughs, given their random nature, we discarded the search for generally applicable algorithms that may systematically identify such cases within a short time-span.

Redner's algorithm as a benchmark

We further implemented the algorithm to identify breakthrough papers developed by Redner (2005) and use it as benchmark for our algorithms. We applied this algorithm to all papers of types article and letter from the period 1990–1994. No restrictions on the size of the citation-windows or the minimum number of received citations were imposed.

3.2. Robustness of the algorithms

We define 'robustness' of an algorithm as the ability to identify the same breakout paper(s) irrespective of the total number of citations a paper received within two years. We tested the robustness empirically by implementing citation count thresholds of 1, 2, 4, ..., 1024 citations. Table 2 shows the performance and the robustness of the algorithms for different thresholds values (2 ... 512) when applied to the Categories and Clusters datasets; for threshold values ≥ 512 the number of publications in the datasets soon drops to 0. As a point of reference, the results for Redner's algorithm (Redner, 2005) are also shown. Table 3 shows the overlap in results between Redner's algorithm and each of our algorithms when no threshold applied.

The RII and CDI detection algorithms manage to capture many breakouts and are most effective for both datasets (Categories and Clusters) because these algorithms focus on the

Table 2: Number of papers in the Categories and Clusters datasets after applying a threshold value for the number of references a publication received within the first 24 months and per algorithm the number of selected publications as function of the applied threshold (2, 8, 32, 128, 512)

Categories		ARI	CDI_{sc}	RII	DII	RNI	Redner's algorithm
Threshold value	Dataset size			Number of documents selected			
≥ 2	252,316	264	1,276	3,543	576	19	6,150
≥ 8	156,765	264	1,276	3,543	74	19	5,883
≥ 32	13,583	36	1,246	3,543	0	14	3,748
≥ 128	539	0	375	539	0	4	511
≥ 512	7	0	7	7	0	0	7
Clusters		ARI	CDI_{dc}	RII	DII	RNI	Redner's algorithm
Threshold value	Dataset size			Number of documents selected			
≥ 2	214,119	60	13,477	3,501	673	8	6,311
≥ 8	137,969	60	13,451	3,501	74	8	5,839
≥ 32	13,369	56	6,930	3,501	0	7	3,748
≥ 128	534	0	486	534	0	4	508
≥ 512	7	0	7	7	0	0	7

Table 3: Overlap of the results of Redner's algorithm and the five algorithms (ARI, DII, RII, CDI, RNI) (no threshold applied)

	Categories	**Clusters**
Algorithms	*Number of papers marked as breakout*	
Redner	6,150	6,311
Redner ∩ ARI	8	11
Redner ∩ CDI	943	3,210
Redner ∩ RII	2,119	2,108
Redner ∩ DII	0	0
Redner ∩ RNI	13	6

more frequently occurring of discovery type 'Charge'. The CDI rates are much higher in Clusters because the CWTS document-clustering method groups documents together on the basis of citation relations. These document-clusters may contain papers from multiple WoS subject categories; this means that diversity is in fact already achieved within a cluster, thereby reducing inter-cluster relations. The consequence is that a different and lower threshold level is used to select breakouts when CDI is applied to Clusters. In the long run, this 'vanishing diversity' effect largely disappears.

ARI and especially RNI are much more targeted towards rarer types of breakouts, because ARI focuses on breakouts that bridge the gap between discovery-oriented science and more application-oriented science. The focus of RNI is on areas where the fabric of the citation network is denser. DII sits between these extremes but is by far the most threshold-sensitive

algorithm, within both Categories and Clusters; it ceases to be effective above the threshold of 16 citations. By virtue of their search criteria, DII and RNI work best within social networks and micro research areas with low-citation levels.

In contrast, the breakout hit rate of RII is only affected by higher ($\geq$ 64) values of the threshold, which follows directly from the requirement that in order to be selected as a breakout paper, it has to be cited by at least 52 papers within 24 months; this high threshold for RII is explained above. The performance of RNI is only slightly threshold sensitive until a threshold of 64 citations. RNI is a very selective algorithm as it searches for sets of citing papers with relatively large numbers of cross-citation relationships. The results are identical for Categories and Cluster and decrease above the threshold of 64 citations within 2 years. ARI selects four times more papers in Categories than in Clusters. A possible explanation is presented in the subsection 'The ARI anomaly' on (page 28).

CDI-generated hit rates are significantly affected within Clusters, although the number of identified breakouts remains large, because of the already discussed way the datasets are constructed. In all, RII is robust up to a threshold value of 32 citations, and for CDI and RNI the robustness starts to break down at a threshold value of 16 citations. Beyond this threshold value the hit rates start to decrease. For ARI this hit-rate breakdown starts for Categories at a threshold value of 8, but for Clusters at 16 — the same value as for CDI and RNI. The DII algorithm should be considered not to be robust, as its hit rates already start to decrease at a threshold value of 2 citations.

As an indication of the performance of the algorithms, we calculated on the basis of both datasets for each algorithm the number of papers recognized uniquely by an algorithm as well as the number of papers recognized by multiple algorithms. We observe (Table 4) that, except for RII, the performance of the algorithms varies for the datasets when measured in absolute numbers of breakout papers. This table also shows the ability of each of the algorithms, regardless of the dataset, to select papers that are not selected by any of the other algorithms that we developed. Because papers can be selected by multiple algorithms the total count is not an add-up of the counts for the individual algorithms.

An increasing threshold for the number of citations a paper received within 24 months results in a decreasing number of papers in a dataset. Furthermore, it is expected that with an increasing threshold, each of the algorithms selects less papers. Each algorithm shows a different response on the increasing threshold levels. Both CDI and RII select above a threshold value of 128 almost all documents and reach the 100%-level for higher threshold values.

Table 4: Performance of the algorithms on the two datasets

	Number of breakout papers identified	of which matched by one algorithm	of which also matched by one or more of the other algorithms
Categories			
Total	*4,946*		
ARI	264	99,6%	0,4%
CDI	1,276	21,2%	78,8%
RII	3,544	71,4%	28,6%
DII	577	99,8%	0,2%
RNI	19	31,6%	68,4%
Clusters			
Total	*15,074*		
ARI	60	50,0%	50,0%
CDI	13,477	78,9%	21,1%
RII	3,544	20,8%	79,2%
DII	674	100,0%	0,0%
RNI	8	12,5%	87,5%

3.3. But is it a potential breakthrough?

As explained, there is no objective measure to qualify or classify a scientific discovery, or its underpinning papers, as a breakthrough. Concepts or criteria from information science cannot be used because there is no straightforward or transparent heuristics for decision-making. One has to rely on assessments based on expert opinion and therefore accept a degree of subjectivity. Various assessment methods, each with relatively high levels of inter-rater reliability, offer guidance. The following additional verification metrics were used:

Scholarly publications supporting Nobel Prizes

If a Nobel Prize in physics, chemistry or physiology or medicine is awarded for a single discovery or invention it considered a 'breakthrough'. The single publication or group of closely related publications in which such a discovery is presented signal this breakthrough. We found eight awarded Nobel Prizes where scholarly work published between 1990 and 1994 was seen by the Nobel Prize committee as being of seminal importance. Five of those cases involve at least one of our identified breakout papers, now verified as a 'breakthrough' paper;

Nature's 'Top-100 list of papers most cited ever'

The papers appearing on Nature's 'Top-100 list of papers most cited ever' (van Noorden et al., 2014) are considered by the scientific community of particular importance. Not all papers on this list display breakthroughs in science by definition as is mentioned in one of the comments to this list Padhi (2014), but experts are able to judge. Thirteen

of our breakouts occur on this list. Two of these papers Laskowski et al. (1993) and Moncada et al. (1991) are not included in our tests because of their document type: 'software review paper' respectively 'review paper', which were excluded from our analysis.

Citations from review papers, patents or social media

We apply three additional methods to help verify our identified breakout papers — all are based again on citation impact, but now these citations are from sources other than 'articles' and 'letters': review papers, patents and social media.

1. Number of times a paper is cited in WoS-indexed review papers. Review papers provide an overview of the developments that occurred in a topical field of science over a certain period of time. Publications that are highly cited by review papers are seen to be important for the developments in a field of science;
2. Number of times a paper is cited in patents. Scholarly papers cited in patents bare relevance to the invention described in the patent and are part of the scientific basis for the developments in a field of technology. These citations link the two domains 'science' and 'technology'. Only a small number ($\approx$ 6%) of the scholarly papers are cited in patents. Based on the number of times cited by patents 11 out of the 60 papers in the test set belong to the top 2% percentile;
3. Number of times a paper is cited in worldwide social media (2012–2014). We conclude that a breakout paper stands out when it is still cited in social media 20^{+} years after publication (1990–1994). Such scholarly papers should be at least looked at to see if they are really special.

For reasons of resource constraints the verification for these three additional methods could not be applied to the full set of breakouts but was done within a small sample of breakout papers *the 60-paper test set*. This test set was constructed by applying the five algorithms to the two datasets Clusters and Categories separate. From each of these 10 applications, we selected the top-10 most cited papers in terms of citation count frequencies. This test set included 60 unique papers (40 of the 100 papers occurred more than once), of which 25 occur both in Categories and in Clusters, 20 exclusively in Categories and 15 are found only in Clusters.

The test results highlight the ability of the RII and CDI algorithms to identify Nature's Top 100 most-cited publications. More importantly, all these breakouts were also cited in at least one review paper. Patents also cite more than half of all breakouts detected by the RII, CDI and RNI algorithms, thus giving an indicator of the technological impact of the scientific discovery. These three algorithms also captured breakouts that generate, or still generate,

Table 5: Percentage of papers that belong to the top 3% percentile for all papers (articles and letters) based on the number of citations received from the different sources for breakout papers and non-breakout papers in the test set

	Cited by review papers		**Cited by patents**		**Times cited within 24 months**
	Breakout papers	Non-breakout papers	Breakout papers	Non-breakout papers	Breakout papers
ARI	93%	64%	7%	27%	100%
CDI	100%	39%	41%	0%	100%
RII	100%	50%	42%	6%	100%
DII	6%	95%	0%	30%	6%
RNI	92%	67%	17%	23%	100%

a wider societal impact, when measured on the basis of social media ('altmetrics') for the years 2012–2014. The CWTS' social media database contains social-media data related to Internet blogs, news, Twitter and Facebook messages collected from the altmetric database provider Altmetric.com[34]. The two 'large-output' algorithms RII and CDI manage to produce the largest number of verified breakouts.

Applying each of the algorithms to the test set results in two groups of documents for each algorithm. One group contains the papers that are selected (breakout papers) and the other group the papers not selected (non-breakout papers). To search for differences in the characteristics of the documents in both groups, the share of papers belonging to the top 3% percentile is used. As the often-used top 10% percentile did not show differences in behaviour between breakout and non-breakout papers, we chose to use the top 3% percentile. These top 3% percentiles are based on the distribution of the number of citations received from the different sources by all papers (letters and articles) published in 1990–1994 that are covered in the WoS database; Table 5 shows the results.

3.4. Breakout classification

We classified the results of the algorithms across the following two dimensions (1) the number of times cited by review papers and (2) the number of times cited by patents. The distribution of papers receiving a certain number of citations within a period of time is highly skewed. Because of this skewness we classified documents in the four classes 'Top 1%' = [99–100]%, 'Top 5%' = [95–99)%, 'Top 10%' = [90–95)% and '<Top 10%' = [0–90)%. Nearly 60% of all papers cited within 24 months are classified on both dimensions 'Citations by review papers' and 'Citations by patents' as '<Top 10%'. For all publications (articles + letters) from 1990–1994 their share exceeds the 78% mark.

[34] http://www.altmetric.com

Applying the algorithms increases the shares of papers in the Top 10% percentiles. Based on this analysis we conclude that by further zooming in on the Top 10% percentile papers the breakout papers can be classified as:

Breakthrough: publications that are part of the scientific basis of Nobel Prize awarded discoveries.

Breakthrough by proxy: publications that belong to the top 1% percentile on the basis of the number of citations from review publications and at the same time to the 1% percentile of the number of citations from patents. These are the publication in the Top 1% row and at the same time in the Top 1% column.

Science-oriented breakthrough by proxy: publications that belong to the top 1% based on the number of citations from review publications but are not significantly cited from patents. These are the publications in the Top 1% row.

Technology-oriented breakthrough by proxy: publications that are not particularly highly cited by review publications but are in the top 1% based on citations from patents. These are the publications in the Top 1% column.

Breakout: a publication identified by at least one of the algorithms that does not belong to one of the four types defined above, but nevertheless worthwhile to take a look at.

Non-breakout: a paper not selected by any of the algorithms and therefore most likely not a (potential) breakthrough

4. Discussion

This developmental study, suffering from inevitable constraints in terms of time and available resources, left several open questions and unresolved problems that were not (sufficiently) addressed and therefore open for further discussion and follow-up work. This section reflects on those topics.

4.1. Research questions

Can the algorithms be used as a generally applicable method to identify breakout papers and if so under what data availability conditions?. Although the algorithms are designed for the early stage identification of discoveries in science represented by research publications — as indexed by Clarivate Analytics *Web of Science* database (WoS) — the detection algorithms can be applied without alterations to other databases that (1) contain bibliographic data of scholarly publications, (2) provide citation relations that interlink publications, (3) contain time stamps to enable systematic tracking and monitoring of temporal developments, and

(4) provide a representative picture of scientific research over time in all relevant fields of science.

What are the similarities and differences between the algorithms in terms of their ability to detect breakouts?. All our detection algorithms are able to identify breakout papers; the resulting datasets show overlap and differences. Some of the breakout papers also stand out in citations given in patents and review papers[35], and are cited by social media sources. The five algorithms can be divided into three groups based on the breakout-detection specificity (recall rate). Group 1 consists of the CDI and RII algorithms. For these algorithms the recall rate increases with increasing threshold values and reaching above a certain threshold value (64 for RII, and 128 for CDI) a situation in which the algorithm selects all remaining papers. The second group consists of ARI and RNI. These algorithms also show an increasing recall rate, but above a certain threshold value (32 for ARI, and 128 for RNI) they break down and fail to select any documents. DII forms a group by itself as the recall rate continuously decreases with increasing threshold values.

Redner's algorithm (Redner, 2005) can be considered a high-performance algorithm and therefore falls in group 1 together with our CDI and RII algorithms. For threshold values from 256 citations and above, the algorithms in this group select all papers remaining in the dataset as breakout paper. Redner's algorithm selects 6,150 papers in Categories and 6,311 in Clusters; 5,907 of these papers belong to both datasets, 243 belong only to Categories, and 404 only to Clusters. From this we conclude that the performance of Redner's algorithm is largely independent of the dataset to which the algorithm is applied. This result is expected, as 'disciplines' are not addressed in Redner's algorithm, and therefore the differences are caused by differences in the contents of the datasets.

The outcomes of the robustness calculations show that the algorithms CDI, RII, and RNI are the ones that — up to a threshold value of 32 citations — are almost unaffected by the value of the threshold. The DII algorithm is the most sensitive of our five algorithms for thresholds imposed on the data. The behaviour of the DII algorithm is different as it focuses on research where the discovery involves a paradigm-shift that starts within a small group of researchers; the core group. Given the short measuring period after publication, the probability for a publication describing this discovery to get cited by authors outside of the core group is limited; therefore, the number of papers selected by DII shows a sharp decrease for larger threshold values.

Both ARI and CDI perform different when applied to Categories and to Clusters; this is not the case for RII, DII, and RNI. The fact that the CDI algorithm behaves differently on

[35] After 20^+ years belonging to the ones highly cited by review papers or by patents

both datasets is to be expected. In 12 cases the CDI_{dc} algorithm when applied to the test set selected a paper, whereas CDI_{sc} did not; the situation that a paper was selected by CDI_{sc} and not by CDI_{dc} dit not occur. This difference in behaviour is caused by the different definitions for 'discipline' used in both datasets.

As 'Charge' breakouts are the more common variant, because there is no change in the theoretical framework or paradigm shift involved, it comes as no surprise that RII and CDI are the algorithms that select the most papers as a breakout. There is no 'overall winner' among the algorithms due to the fact that each is developed with a particular type of breakthrough in mind. The definitive conclusion that a breakout paper really presents a breakthrough must be based on information other than bibliographic information.

Can we determine the effectiveness of each algorithm in terms of identifying breakout papers that are generally regarded as breakouts and potential breakthroughs?. The combination of the five algorithms identified all 11 papers of WoS-type 'article' or 'letter' published in the period 1990–1994 that occurred in Nature's 'Top-100 list of most cited papers ever'. For five of the eight Nobel Prizes in Chemistry, Physics, and Physiology or Medicine for which scholarly work published between 1990 and 1994 forms the scientific basis, at least one of the founding papers was detected.

Redner's algorithm selects more breakout papers[36] than our algorithms (Table 2) and the results partially overlap (Table 3). The only exception is the CDI algorithm applied to the clusters dataset. The method proposed by Redner takes into account all citations, whereas the algorithms we developed focus on the citation dynamics of a paper within 24–36 months[37] after publication. Redner's algorithm identifies in total 36 of the 60 papers in the test set. Except in the case of the DII algorithm there is overlap between the results of our algorithms and Redner's algorithm, sometimes a very small one.

Except for the (DII) algorithm the selected papers are 'high' or 'very high' cited by review papers, they are cited in patents, and received citations within 24 months.

The breakthrough publications that form the basis of the four case studies (Winnink and Tijssen, 2014, 2015; Winnink et al., 2013, 2016) received from review articles within 24 months at least 4 citations, and until the beginning of 2016 at least 73. Of the papers in the validation test-set 32 belong to the top 1% percentile based on the citations from review papers received within the first 24 months after publication. These 32 publications in the top 1% percentile after 24 months are also in the top 1% in the beginning of 2016;

[36] In our opinion is what we call a 'breakout' identical to what Redner calls a 'breakthrough'

[37] This period of 24–36 months was chosen in order to stay as close as possible to the moment of publication. Other time periods are possible, e.g. Rogers (2010) uses a five years windows and Ponomarev et al. (2014) use a two-step forecasting model that combines short citation periods (of 6, 12 or 24 months) where highly-cited publications are monitored for periods up to 5 years

this observation is in line with (Adams, 2005). We classify publications that are highly referenced by review articles — belong to the top 1% percentile — as potential breakthroughs: 'Breakthrough by proxy' or 'Science-oriented breakthrough by proxy'

4.2. Limitations of this study

Does the use of a time window of 24–36 months cause some breakouts to be inadvertently not recognised?.

By focusing on the first 24–36 months after publication of a paper we ignore 'sleeping beauties' (van Raan, 2004, 2015). We also did not address the situation in which the citation profile of a paper at early stage gives the impression that it presents a 'breakthrough' that later turns out not to be the case.[38]

Preliminary results of a small follow-up study shows that for almost 92% of the publications that show 'breakout character' during the first 10 years after publication this behaviour manifests itself in the first year; therefore focussing on the time period of 24–36 months with the publication date as point of reference seems to be appropriate. Changing the algorithms so the search for breakouts not only starts at the moment of publication but also one year after increases the hit rate with an extra 6.4%.

The ARI *anomaly.*

The performance of ARI on the two datasets (Clusters and Categories) differs; it detects four times as many breakout papers in Categories as it does in Clusters. ARI searches for papers that are supposed to act as bridges between discovery-oriented science and application-oriented science. Approximately 36% of the papers in the source data set (WoS) are characterised as discovery-science oriented. The share of discovery-science oriented papers is above this average for Clusters and equal to this average for Categories. The fact that ARI selects more breakout in Categories than in Clusters seems counter-intuitive as the datasets are constructed from the same data source by conceptually equivalent methods. The factors we believe that play a role in this 'ARI anomaly' are:

1. The document selection process distributes the papers among 823 clusters (out of 865), and among 199 categories (out of 251). This results in an average of 106 discovery-science papers per cluster, and 461 per category;
2. For 60% of the 199 categories, the share of discovery-science papers is above the overall average of 36%; for the 823 clusters this share is equal to the overall average;
3. Discovery-science papers receive on average 4.7 citations within 24 months compared to 3.8 for application-science papers;

[38] An example is Fleischmann and Pons (1989) in which the existence of nuclear fusion at room temperature — 'cold fusion' — is claimed. This claim was almost immediately criticized but it was not before 2012 that in Dmitriyeva et al. (2012) the definitive conclusion that the claim was false was drawn

4. Papers can have more than one subject category — on average 1.5 — assigned to them but can be a member of only one document cluster. Therefore, the same paper might be selected multiple times (for different subject categories) during the selection of documents for Categories. The selection method creates a bias towards highly cited papers to which multiple categories are assigned and in this way preventing other less cited papers to be selected;
5. On average fewer subject categories are assigned to discovery-science papers than to the more applied-science oriented papers.

In our opinion, these factors in combination with the method of constructing the two data sets causes higher cited discovery-science papers to be preferred in the selection of papers for Categories, and thereby account for the higher performance of ARI.

Is there an (implicit) link in the algorithms between science and technology?.
The algorithms developed in this study are constructed on the basis of the outcomes of case studies. One of the criteria used to select cases was that the scientific breakthrough discoveries resulted in new technological developments, as shown by the occurrence of citations from patents. In this way the algorithms may contain in an implicit form a link between science and technology that could explain the occurrence of patent citations.

Retracted publications.
The retraction of scientific publications is increasing; the number of retracted papers in MEDLINE®[39] reached the 1% level in 2006 (Cokol et al., 2008). The mean time to retract a publication depends on the reason to retract and ranges from 26 to almost 47 months (Steen et al., 2013, Table 1). Retracted publications do not vanish from the scientific knowledge base and are still cited even after their retraction (van Noorden, 2011); in only 8% of the citations the retraction is mentioned. Retracted articles live on in personal libraries and on the Internet (Davis, 2012). Retracted publications are therefore in general present as a referenced publication or as a citing document in the first 24–36 month after publication period that is used in this study. After the identification of breakout papers a check for retractions should be carried out to prevent such papers to be seen as a potential breakthrough.

4.3. Options for further research

We see at least the following options for further research on in order to improve and expand the analytical framework presented in this paper:

- Further refinement of the algorithms, e.g. influence of parameter values on an algorithms‘ performance;

[39] MEDLINE® is the U.S. National Library of Medicine (NLM) premier bibliographic database that contains more than 22 million references to journal articles in life sciences with a concentration on biomedicine

- It is assumed that the algorithms are time invariant; this is however not further investigated in this study. This point is partly addressed in two preliminary follow-up studies that use publications from the period 2007–2011; these studies did not show significant different results. A systematic study should be carried out to resolve this issue'
- This study shows that the results of the algorithms show overlap. This interdependency of the algorithms should be further investigated as this might result in insights in the factors that play a role in the 'forming' of discoveries and especially of breakthroughs;
- Construction of new algorithms for early stage identification of breakout papers;
- *Sliding window* versions of the algorithms to analyse a paper's breakout character over time;
- The algorithms focus on different types of discoveries in Kosland-sense. The framework also offers a method to classify scientific publications on the basis of their breakout character. This classification is presented on page 25. Further research is needed to find out if these two classifications can be used to analyse the progress of science.
- The remarkable observation that RNI, our 'lowest-output' algorithm, has such a high hit rate in terms of selecting breakout papers raises the question 'How effective is the RNI algorithm in detecting breakthroughs?' Further in-depth research is needed to answer this question as no definitive conclusion can be given on the basis of the available bibliographic information.
- The implemented framework facilitates generating datasets that consist exclusively of scientific discoveries that are considered to have an above average impact on science. Such 'clean' datasets can be used for, large scale, analysis of the dynamics of the science system from the perspective of high-impact discoveries. Especially the search for general mechanisms that stimulate the emergence of discoveries in science, in particular breakthroughs, could benefit from such clean datasets.

5. Conclusions

The way the scientific community reacts on a discovery determines if it is to be considered a 'breakthrough'. This reaction of the scientific community is reflected in bibliographic time dependent signals. Guided by general characteristics of a breakthrough — a suddenly occurring event that has a major impact on follow-up scientific research — we searched for characteristic patterns in the citation profiles of known breakthrough discoveries. Five algorithms each focusing on different aspects of citation profiles of individual publications were developed and implemented. These algorithms focus on the citation profiles during the

24–36 months after publication. The algorithms classify publications in five types ranging from 'non-breakout' to 'breakthrough by proxy'. For a definitive conclusion on the sixth type — 'breakthrough' — the algorithms cannot decide at early stage as additional information, particularly expert opinions are indispensable. It is argued that the early-stage characterisation by our algorithms provides is a reliable measure of a papers long-term impact on science.

The aim of this study was to develop an analytical framework and measurement model consisting of general applicable algorithms to capture the dynamics of the diffusion of scholarly knowledge and conclude at early stage if a paper should be considered a breakout. We succeeded in detecting many breakout papers with distinctive impact profiles. A small subset of these breakouts is classified as 'breakthroughs': Nobel Prize research papers; papers occurring in Nature's Top-100 Most Cited Papers Ever; papers still (highly) cited by review papers or patents; or those frequently mentioned in today's social media. We also compare the outcomes of our algorithms with the results of a 'baseline' detection algorithm developed by Redner in 2005, which selects the world's most highly cited 'hot papers'.

The analytical framework presented can be seen as an operational, probably incomplete, definition of a breakthrough. We conclude that "It is possible to design, develop, implement and test an analytical framework and measurement model as a general-purpose tool that uses bibliographic information for early detection of potential breakthroughs in science".

Uncertainty is an integral part of data that comes from observations. As the data sets increase in size more precise answers can be derived while on the other hand the chances of false findings increase exponentially (Spiegelhalter, 2014). Spiegelhalter argues that in order to avoid such false findings statistical analysis of large data sets (Big Data) should be accompanied by knowledge of the limitations and strengths of the models that are taken into account. The algorithms developed in this study can help in preventing false findings when searching for potential breakthrough publications by analysing large bibliographic datasets.

Acknowledgments

We thank Professor Redner for answering questions on details of his algorithm and our colleague Ludo Waltman for his help in applying the CWTS publication-clustering method to produce the dataset Clusters. We also like to thank David Pendlebury for providing insight into the method used by Thomson Reuters to identify *Nobel prize class* papers. Furthermore, we like to thank the editor and the reviewers for their valuable comments on the first version of this paper.

References

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, 63:567–581. 10.1007/s11192-005-0228-9.

Andersen, H., Barker, P., and Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge University Press, New York, NY, USA.

Ball, P. (2004). *Critical Mass - How One Thing Leads to Another*. Arrow Books, London.

Baumgartner, S. E. and Leydesdorff, L. (2014). Group-based trajectory modeling (gbtm) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims". *Journal of the Association for Information Science and Technology*, 65(4):797–811.

Becattini, F., Chatterjee, A., Fortunato, S., Pan, R., Parolo, P., and Mitrovic, M. (2014). *The Nobel Prize* (http://scitation.aip.org/content/aip/magazine/physicstoday/news/10.1063/pt.5.2012).

Bettencourt, L. M. A., Kaiser, D., Kaur, J., Castillo-Chávez, C., and Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3):495–518.

Bettencourt, L. M. A. and Kaiser, D. I. (2011). General critical properties of the dynamics of scientific discovery. Technical report, Office of Scientific and Technical Information - U.S. Department of Energy.

Bettencourt, L. M. A. and Kaiser, D. I. (2015). Formation of scientific fields as a universal topological transition. *arXiv*, 1504.00319(v1):1–8.

Bettencourt, L. M. A., Kaiser, D. I., and Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210–221. Science of Science: Conceptualizations and Models of Science.

Bonaccorsi, A. (2008). Search regimes and the industrial dynamics of science. *Minerva*, 46:285–315.

Bonaccorsi, A. (2010). New forms of complementarity in science. *Minerva*, 48:355–387.

Brannigan, A. and Wanner, R. A. (1983). Historical distributions of multiple discoveries and theories of scientific change. *Social Studies of Science*, 13(3):417–435.

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics*, 12(5-6):373–379.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., and Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3):191–209.

Ciechanover, A. (2009). Tracing the history of the ubiquitin proteolytic system: The pioneering article. *Biochemical and Biophysical Research Communications*, 387(1):1 – 10.

Cintron-Arias, A., Bettencourt, L. M. A., Kaiser, D. I., and Castillo-Chávez, C. (2005). On the transmission dynamics of knowledge. Technical report, Office of Scientific and Technical Information - U.S. Department of Energy.

Cokol, M., Ozbay, F., and Rodriguez-Esteban, R. (2008). Retraction rates are on the rise. *EMBO Reports*, 9(1):2–2.

Crane, D. (1972). *Invisible colleges: diffusion of knowledge in scientific communities*. The University of Chicago Press, Chicago.

Cyranoski, D. (2004). Korea's stem-cell stars dogged by suspicion of ethical breach. *Nature*, 429(6987):3–3.

Darden, L. and Maull, N. (1977). Interfield theories. *Philosophy of Science*, 44(1):43–64.

Davis, P. M. (2012). The persistence of error: a study of retracted articles on the internet and in personal libraries. *Journal of the Medical Library Association : JMLA*, 100(3):184–189.

Dmitriyeva, O., Cantwell, R., McConnell, M., and Moddel, G. (2012). Origin of excess heat generated during loading Pd-impregnated alumina powder with deuterium and hydrogen. *Thermochimica Acta*, 543:260 – 266.

Fleischmann, M. and Pons, S. (1989). Electrochemically induced nuclear-fusion of deuterium. *Journal of electroanalytical chemistry*, 261(2A):301–308.

Gerstein, M. and Douglas, S. (2007). RNAi development. *PLoS Computational Biology*, 3(4):e80–775.

Globe, S., Levy, G. W., and Schwartz, C. M. (1973). Science, technology and innovation. NSF-study NSF-C-667, Battelle Memorial Inst., Columbus, Ohio, Columbus Labs.

Green, S. J. and Brendsel, J. (2008). Letter to the editor: Key discoveries often originate with lone researchers. *Nature*, 456:315.

Grupp, H. and Schmoch, U. (1992). *Dynamics of Science-Based Innovation*, chapter 9 At the crossroads in laser medicine and polyimide chemistry: patent assessment of the expansion of knowledge, pages 269–301. Springer-Verlag.

Guimerà, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702.

Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C., and Miller, M. D. (2000). Inhibitors of strand transfer that prevent integration and inhibit hiv-1 replication in cells. *Science*, 287(5453):646–650.

Heilbron, J. L. (1972). Book review: Illinois Institute of Technology Research - Technology in Retrospect and Critical Events in Science. *Isis*, 63(1):115.

Hofstadter, D. R. and Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books, New York.

Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. In Hannaway, C., editor, *Biomedicine in the Twentieth Century: Practices, Policies, and Politics*, volume 72 of *Biomedical and Health Research*, pages 317–353. National Institutes of Health, Bethesda, MD.

IIT Research Institute (1968). Technology in retrospect and critical events in science. vol. 1. Technical report, IIT Research Institute; National Science Foundation.

IIT Research Institute (1969). Technology in retrospect and critical events in science. vol. 2. Technical report, IIT Research Institute; National Science Foundation.

Isenson, R. S. (1969). Project hindsight (final report). Technical Report AD495905, Office of the Director of Defense Research Engineering, Washington, DC, 20301.
Jewkes, J., Sawers, D., and Stillerman, R. (1958). *The Sources of Invention*. MacMillan, London.
Jones, B. F., Wuchty, S., and Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262.
Julius, M., Berkoff, C. E., Strack, A. E., Krasovec, F., and Bender, A. D. (1977). A very early warning system for the rapid identification and transfer of new technology. *Journal of the American Society for Information Science*, 28(3):170–174.
Koshland, D. E. (2007). The cha-cha-cha theory of scientific discovery. *Science*, 317(5839):761–762.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. The University of Chicago Press, Chicago.
Lade, S. J. and Gross, T. (2012). Early warning signals for critical transitions: A generalized modeling approach. *PLoS Comput Biol*, 8(2):e1002360.
Laskowski, R., MacArthur, M., Moss, D., and Thornton, J. (1993). Procheck - a program to check the stereochemical quality of protein structures. *Journal Of Applied Crystallography*, 26(2):283–291.
Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceedings of the ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, number ACM 1-59593-135-X/05/0008 in KDD '05, pages 177–187, New York, NY, USA. ACM.
Lievrouw, L. A. (1989). The invisible college reconsidered: Bibliometrics and the development of scientific communication theory. *Communication research*, 16(5):615–628.
Mandelbrot, B. B. (1982). *The fractal geometry of nature*. W.H. Freeman and Company, New York.
McGrew, J. F., Bilotta, J. G., and Deeney, J. M. (1999). Software team formation and decay: Extending the standard model for small groups. *Small Group Research*, 30(2):209–234.
Meeussen, L., Delvaux, E., and Phalet, K. (2014). Becoming a group: Value convergence and emergent work group identities. *British Journal of Social Psychology*, 53(2):235–248.
Merton, R. K. (1961). Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470–486.
Merton, R. K. (1965). *On the shoulders of giants - A Shandean Postscript*. The University of Chicago Press, Chicago, Ill., USA.
Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.
Moed, H. F., Glänzel, W., and Schmoch, U., editors (2004). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
Moncada, S., Palmer, R., and Higgs, E. (1991). Nitric-oxide - physiology, pathophysiology, and pharmacology. *Pharmacological Reviews*, 43(2):109–142.
Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. V., Grigorieva, I. V., and Firsov, A. A. (2004). Electric field effect in atomically thin carbon films. *Science*, 306(5696):666–9.
Ogburn, W. F. and Thomas, D. (1922). Are inventions inevitable? A note on social evolution. *Political Science Quarterly*, 37(1):83–98.
Padhi, B. (2014). Comment on 'The Top 100 Papers - Nature explores the most-cited research of all time'.
Perla, R. J. and Carifio, J. (2005). The nature of scientific revolutions from the vantage point of chaos theory. *Science & Education*, 14:263–290.
Ponomarev, I. V., Williams, D., Lawton, B., Cross, D. H., Seger, Y., Schnell, J., and Haak, L. (2012). Breakthrough paper indicator: early detection and measurement of ground-breaking research. In Jeffery, K. G. and Dvořák, J., editors, *E- Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems*, pages 295–304.
Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., and Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81(January 2014):49–55.
Price, Derek J. de Solla. (1963). *Little science, big science*. Columbia University Press.
Price, Derek J. de Solla. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.
Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25:358–359.
Redner, S. (2005). Citation Statistics from 110 Years of Physical Review. *Physics Today*, 58(6):49–54.
Rogers, J. D. (2010). Citation analysis of nanotechnology at the field level: implications of R&D evaluation. *Research Evaluation*, 19(4):281–290.
Scheffer, M. (2009). *Critical Transitions in Nature and Society*. Princeton Studies in Complexity. Princeton University Press, Princeton, NJ, USA.
Scheffer, M. (2010). Complex systems: Foreseeing tipping points. *Nature*, 467(7314):411–412.
Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260):53–59.
Schneider, J. W. and Costas, R. (2017). Identifying Potential "Breakthrough" Publications Using Refined Citation Analyses: Three Related Explorative Approaches. *Journal of the Association for Information Science and Technology*, 68(3):709–723.
Simonton, D. K. (1978). Independent discovery in science and technology: A closer look at the poisson distribution. *Social Studies of Science*, 8(4):521–532.

Simonton, D. K. (1979). Multiple discovery and invention: Zeitgeist, genius, or chance?. *Journal of Personality and Social Psychology*, 37(9):1603–1616.
Skilton, P. (2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics*, 78(3):525–542.
Small, H. G. (1977). Co-citation model of a scientific specialty - longitudinal-study of collagen research. *Social Studies of Science*, 7(2):139–166.
Small, H. G., Boyack, K. W., and Klavans, R. (2013). Identifying emerging topics by combining direct citation and co-citation. In Gorraiz, J., Schiebel, E., Gumpenberger, C., and Moed, H. F., editors, *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference 15–17 July Vienna, Austria*, volume I, pages 928–940.
Spiegelhalter, D. J. (2014). STATISTICS The future lies in uncertainty. *SCIENCE*, 345(6194):264–265.
Steen, R. G., Casadevall, A., and Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE*, 8(7):1–9.
Sung, J. (2008). Embodied anomaly resolution in molecular genetics: A case study of RNAi. *Foundations of Science*, 13(2):177–193.
Tijssen, R. J. W. (2010). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology*, 61(9):1842–1852.
Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, 63(6):384–399.
Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.
Uzzi, B. and Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447–504.
van Andel, P. (1994). Anatomy of the unsought finding. serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science*, 45(2):631–648.
van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature*, 478(7367):26–28.
van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers - nature explores the most-cited research of all time. *Nature*, 514(7524):550–553.
van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, 59(3):467–472.
van Raan, A. F. J. (2015). Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLoS ONE*, 10(e0139786):1–38.
Vitanov, N. and Ausloos, M. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion. In Scharnhorst, A., Börner, K., and van den Besselaar, P., editors, *Models of Science Dynamics*, volume 69 of *Understanding Complex Systems*, pages 69–125. Springer Berlin / Heidelberg. 10.1007/978-3-642-23068-4_3.
von Bertalanffy, L. (1969). *General System Theory*. George Braziller, Inc, 227 Broadway, New York.
Wallace, P. R. (1947). The band theory of graphite. *Physical Review*, 71(9):622–634.
Walsh, J. (1973). Project review: Technological innovation: New study sponsored by NSF takes socioeconomic, managerial factors into account. *Science*, 1820:846–847.
Waltman, L. and van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392.
Wang, J., Veugelers, R., and Stephan, P. (2016). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Working Paper 22180, National Bureau of Economic Research.
Weinberg, A. M. (1970). Scientific teams and scientific laboratories. *Daedalus*, 99(4):1056–1075.
Whitfield, J. (2008). Collaboration: Group theory. *Nature*, 455(7214):720–723.
Winnink, J. J. (2017). *Early-stage detection of breakthrough-class scientific research: using micro-level citation dynamics*. PhD thesis, Social and Behavioural Sciences, Leiden University.
Winnink, J. J. and 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.
Winnink, J. J. and Tijssen, R. J. W. (2015). Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. *Scientometrics*, 102(1):113–134.
Winnink, J. J., Tijssen, R. J. W., and van Raan, A. F. J. (2013). The discovery of 'introns'; an analysis of the science-technology interface. In Hinze, S. and Lottman, A., editors, *Translational twists and turns: science as a socio-economic endeavor - Proceedings of STI 2013 Berlin (18th International Conference on Science and Innovation Indicators*, pages 427–438.
Winnink, J. J., Tijssen, R. J. W., and van Raan, A. F. J. (2016). Theory-changing breakthroughs in science: the impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology (JASIST)*, 67(5):1210–1223.
Wooding, S. (2007). Setting the historical context for project retrosight. RAND Rand Europe working paper series WR-466-RS, RAND Europe Cambridge Westbrook Centre, Milton Road, Cambridge. CB4 1YG. UK.
Wray, K. B. (2011). *Kuhn's Evolutionary Social Epistemology*. Cambridge University Press, Cambridge, UK; New York.
Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Author cv's

Dr. Jos Winnink.

Jos Winnink is a researcher at the Centre for Science and Technology Studies (CWTS) of Leiden University in the Netherlands. His research focuses on the interaction of science and technology. His research topic Knowledge flows between frontier science and breakthrough technologies is one of the research lines of the Science, Technology, and Innovation Studies (STIS) group. In his PhD Thesis he focused on early stage detection of breakthroughs using bibliometric information. On these subjects he published several (conference) papers. Given his history in the world of patents Jos has a special interest in patent information and is within CWTS one of the experts on this subject. Currently Jos' focus is on linking patent databases with other data sources to create multidimensional data infrastructures.

Prof. dr. Robert Tijssen.

Robert Tijssen holds the Chair of Science and Innovation Studies at Leiden University. He leads the Science, Technology and Innovation (STIS) research program at CWTS. His main interests concern the development and application of analytical frameworks and measurement models to assess trends and patterns within the interdependencies between science, innovation and higher education systems. A major part of his current research agenda are 'innovative universities' and academic entrepreneurs, which is driven in part by activities in university ranking systems and indicator-driven projects: U-Multirank; Leiden Ranking, U21 Ranking of National Higher Education Systems, and European Innovation Scoreboard. This research line connects him to the University of Bologna (Italy) where he holds a visiting fellowship. Robert is also an international research partner at the ESRC-HEFCE funded Centre for Global Higher Education (London, UK). Africa is a major part of his academic profile: he is a research fellow at the LDE Center for Frugal innovation in Africa (Leiden University, Delft University of Technology and Erasmus University Rotterdam). He is also professor at Stellenbosch University (South Africa) and board member of South Africa's DST-NRF Center of Excellence in Scientometrics and Science, Technology and Innovation Policy. Major research topics in this line are research excellence in Africa, and university-supported inclusive innovations. Robert's advisory work, stretching back more than 25 years, covering a portfolio of projects and activities in Europe and elsewhere. Mostly dealing with research management and science/innovation policy issues — both for public sector clients and private companies — the portfolio includes assessments of research programs, advice on quality assurance systems, developing evaluation systems of regional innovation impact, to conducting sector-wide strategic analyses of R&D performance.

Prof. dr. Ton van Raan.

Ton (A.F.J.) van Raan is emeritus professor of Quantitative Studies of Science. Founder and until 2010 Director of the Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands. After his retirement as CWTS Director, he remained research professor. He studied mathematics, physics and astronomy at Utrecht University. PhD in Physics, Utrecht. Post-doctoral fellow at the University of Bielefeld (Germany), visiting scientist in the US, UK, and France. Work in atomic and molecular physics, laser-physics and astrophysics. From 1977 senior research fellow physics in Leiden, in 1985 'field switch' to science and technology studies, 1991 Professor. Research focus: application of bibliometric indicators in research evaluation; science as self-organizing complex system, statistics of bibliometric indicators, ranking and benchmarking of universities, mapping of science. In 1995 he received, together with the American sociologist Robert K. Merton, the Derek de Solla Price Award, the highest international award in the field of quantitative studies of science. He published (as author and co-author) around thirty articles in physics and two hundred in science and technology studies. Prof. van Raan set up a small company Science Consult for advice on research evaluation and science policy issues. On the occasion of his retirement as CWTS director he was awarded by the Queen of the Netherlands with the royal distinction of Knight in the Order of the Dutch Lion. His current research focuses on Sleeping Beauties in Science, scaling properties of universities, and urban scaling.