Early-stage detection of breakthrough-class scientific research : using micro-level citation dynamics

Winnink, J.J.

Chapter 1

Introduction

1.1 General context

Scientific and scholarly research may result in a new discovery: an observation or finding of something unknown prior to that discovery. 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. In this thesis the term *breakthrough* is used for those discoveries that have such a major impact on science. 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.

The impact of discoveries may extend beyond the domain of science. Scientific discoveries may be crucial steps towards technological applications, and to innovations and products. Several well-known studies searched for the impact of those discoveries on the development of technology[1] and have analysed cases of technological progress (e.g. Globe et al. 1973; Heilbron 1972; IIT 1968, 1969; Isenson 1969; Jewkes et al. 1958; Walsh 1973; Wooding 2007). All these studies indicate that it can take many years before a scientific discovery finds its way into new or adapted technology. This observation is also in line with Grupp and Schmoch (1992) who conclude that the

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

diffusion of basic scientific knowledge to become the basis of new technology often takes dozens of years[2] Scientific discoveries and their incorporation in technology are often interlinked in complex ways within research and development (R&D)[3] systems, and may span several years, decades, or even centuries.

Forecasting the changes that discoveries may bring about, and monitoring or nowcasting[4] the evolution of emerging areas in science or technology, presents us with a series of conceptual and methodological challenges. This thesis focuses on how discoveries can 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[5] and innovation[6] 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.

1.1.1 Research goal and research question

This thesis will focus on one of the upstream stages in the process: new scientific discoveries and the first wave of impacts on their scientific environment. At this early stage, it is usually not clear — neither to the researchers responsible for the discovery nor to external analysts — what the implications are (or could be) in terms of scientific impact or usage outside science. Given their restricted view of R&D decision-makers on these new developments, any source of empirical and objective information that may help provide clarity

[2] An example is Graphene. Wallace (1947) presented — based on theoretical physical calculations — the properties to be expected for a material currently known as 'graphene'. It was not until 2004 however, with the publication of 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 For to Konstantin Novoselov and André Geim for this discovery.

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

[4] Nowcasting is the activity of estimating the current situation on the basis of historic data.

[5] 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

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

is highly desirable. To address this information gap, I have developed, tested and applied a series of computerised algorithm-based methods, extracting their information from the vast amount of scientific publications, to identify specific, high-profile discoveries – denoted as 'breakthroughs' – that are expected to have a major impact on future research.

The approach used in this study is to analyse, in detail, four case studies of well-known 'landmark' discoveries that have been classified as a breakthrough by subject experts. We can benefit from a large pool of information to study each of these well documented 'exemplar' discoveries. The bibliographic data related to these discoveries, from the respective research publications, was searched for impact patterns that are characteristic for the early stage evolution of that particular class of breakthroughs. The detection algorithms, designed to single out similar types of breakthroughs, were constructed from those impact patterns.

The objective of this research is to develop and validate methods forming an analytical framework to identify at an early stage high-impact publications that form the basis of new fields in science. An analytical framework should facilitate the nowcasting, and perhaps even the forecasting, of the impact of a scientific discovery on the evolution of science and technology. The founding concepts this analytical framework is based on are (1) a scientific discovery can be considered a discontinuity in the science system as it exists at that moment, (2) and it is possible to detect such discontinuities using bibliographic information. The research goal leads to the following research question:

> *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?*

By focusing on individual scholarly publications this study concentrates on the *micro level* of scientific progress.

1.1.2 Structure of this thesis

This PhD-thesis is structured in the following way. This first chapter continues with sections 1.2–1.5 which introduce (1) the theoretical and conceptual framework of the study, (2) the analytical framework that guides the research design and facilitates data collection, (3) the introduction of the four case studies and the concluding validation study of the detection algorithms, and (4) the information sources used in the study.

The four case studies, which constitute the empirical core of this empirical study, are discussed in full detail in chapters 2–5. Chapter 2 presents the first case study that focuses on the development of 'Integrase Inhibitors', a type of HIV/AIDS medicine. The case study analyses the route from the scholarly publication describing the basic biochemical mechanism (Hazuda et al., 2004b) these medicines try to block up to the granted patent that describes

the first drug of this type that was approved for human use. Chapter 3 discusses the second case study, namely the discovery of a method to isolate Graphene[7], and the effect this discovery (Novoselov et al., 2004) had on the scientific community. The third case study discusses the discovery of the 'Ubiquitin[8] mediated proteolytic system', and is presented in Chapter 4. Particular attention is paid to the fact that this breakthrough was the result of research by a small core team of scholars. Chapter 5 presents case study four that discusses the follow up of the discovery of 'Introns[9]' in 1977 by Phillip Allen Sharp and Richard J. Roberts. The focus in this case study is on the fact that after the first breakthrough discovery a second breakthrough took place. This second breakthrough is marked by a shift from discovery-oriented research to more application-oriented research. Chapter 6 presents the study that validates the five algorithms that arose from the four case studies and can be applied to raw bibliographic data without the need of extensive preprocessing to clean and harmonise the data. Chapter 7 synthesizes the empirical work, introduces an alternative definition for a breakthrough, discusses the results, points to possible directions for future research, presents some preliminary results of my follow-up research, and ends with a final remark.

1.2 Theoretical and conceptual framework

1.2.1 Science as a dynamic system

Science can be considered a dynamic system[10] 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 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 in (von Bertalanffy, 1969). All systems, not just the large ones, can undergo irreversible[11] changes (Mandelbrot, 1982). Several empirical studies have shown that dynamic systems

[7] Graphene is a single, tightly packed layer of carbon atoms that are bonded together in a hexagonal honeycomb lattice. Graphene has a number of extraordinary properties.

[8] Ubiquitin is a label-molecule that is attached to a protein that is to be destroyed, Shortly before the protein is squeezed into the proteasome, its ubiquitin label is disconnected for re-use.

[9] A non-coding, intervening sequence of DNA within a gene

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

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

can transmit early-warning signals indicating a 'phase transition' is about to happen Lade and Gross (2012); Scheffer (2009, 2010); Scheffer et al. (2009); 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.

1.2.2 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"[12]. 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[13] and are expected[14]. 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 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

[12] 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)

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

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

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.

1.2.3 Discoveries in science

Identical or related discoveries frequently come in a manifold - "...It is an interesting phenomenon that many inventions[15] 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 in the following way:

> "...These seeming singletons[16] which then turn out to have been multiples or rediscoveries. Other, more compelling, classes of evidence bear upon the apparently incorrigible hypothesis 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.

1.2.4 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 usually by research articles

[15] 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.

[16] Singletons – discoveries occurring only once.

in scholarly or technical journals ('research publications') or in books. Price (1963, Ch.3, p.68) discusses the role of scientific publications, and concludes:

> "... 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[17] 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[18] 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[19] 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. Notwith-

[17] 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.

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

[19] 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 ...".

standing the limitations, citation relations can be used as a proxy to reveal the 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 multipage 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 multipage 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[20].

1.2.5 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 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 a many reasons.[21] 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)[22] analysed the time lag between discoveries

[20] 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.

[21] 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 (Ciechanover, 2009, p.2)

[22] Becattini et al. (2014) "...After 1985 about 15% of physics, 18% of chemistry, and 9% of medicine prizes were awarded within 10 years of the corresponding discoveries. By contrast,

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 the 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 mass' to undergo a major change. Complex dynamic systems can have 'tipping points' (Scheffer, 2010; Scheffer et al., 2009). 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

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

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 and Kaiser, 2011, 2015; Bettencourt et al., 2009) 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 nonlinear 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[23], 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.

1.2.6 Characterizing discoveries

Time not only picks the winners, it also unmasks discoveries that turned out to be hypes[24], 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 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

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

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

Table 1.1: 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)

in relation to already existing scientific knowledge: ***Cha**rge, **Cha**llenge* and ***Cha**nce.* Koshland's (2007) classification, 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*, in Section 1.3.1 Table 1.1 presents key Cha-Cha-Cha characteristics while Table 1.2 presents some illustrative examples selected by Koshland. Based on the nature of a discovery, events are categorized into one of the three Cha-Cha-Cha types. Discoveries of type Charge are the most common.

1.2.7 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. Three of the discoveries listed in Table 1.2 reflect such 'team work', research where two or more researchers played a key role in the working leading up to the breakthrough. 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

Table 1.2: Examples of discoveries and their classification by Koshland (2007, p.761)

Problem that needed solving	Discovery	Discoverer	Category of discovery
Movement of stars, Earth, and Sun	Gravity	Newton	Charge
Preventing heart attacks	Cholesterol metabolism	Brown & Goldstein	Charge
Why offspring look like their parents	Laws of heredity	Mendel	Charge
Structure of C_6H_6	Benzene structure	Kekulé	Challenge
Constant speed of light	Special relativity	Einstein	Challenge
Atomic spectra that could not be explained	Quantum mechanical atom	Bohr	Challenge
How DNA replicates and passes on coding	Base pairing in double helix	Watson & Crick	Challenge
Clear spots on Petri dish	Penicillin	Fleming	Chance
Preventing heart attacks	Cholesterol metabolism	Brown & Goldstein	Charge
Crystals of D- and L-tartaric acid	Optical activity	Pasteur	Chance
Reagent 'stuck' in storage cylinder	Teflon®	Plunkett	Chance

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[25] 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.

[25] [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)

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 sience, especially since World War II. Such research teams are said to be part of 'big science' (Price, 1963). Workgroups 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. 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.

1.3 Analytical framework

1.3.1 Facing the methodological challenge

Bibliographic information within a research publication that first describes a discovery may refer to its 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[26]. 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*[27]. 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 1.1.1 on page 2 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[28], 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?"

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); Leskovec et al. (2005); Uzzi and Spiro (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 thesis is to tackle this question systematically and in a large-scale 'macro-level' fashion, i.e. scanning world

[26] Metric – a technical system or standard of measurement.

[27] This group is further broken down into subcategories as explained on in Section 6.A.9 on Page 156

[28] Operationalization is the process of strictly defining variables into measurable factors.

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 thesis 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. (2014b) focuses on the effects of interdisciplinarity in the subject categories, and geographical diversity. Schneider and Costas (2015) 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 thesis 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.

1.3.2 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[29]. 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 exemplars' 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[30] changes in publication and citation patterns. In this study, the bibliographic is extracted from the *Web of Science Core Collection* database[31] (abbreviated here to WoS). Further information about this information source, and relevant measurement details, are provided in Section 1.5. 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

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

[30] 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.

[31] Web of Science Core Collection is a product of Thomson Reuters, the USA-based media company. The division of Thomson Reuters responsible for the WoS has recently (July 2016) been sold to two investment firms: Onex Corporation and Baring Private Equity Asia.

(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*[32] (2) in-house defined *document clusters*[33];
6. Application of all the developed detection algorithms to both datasets;
7. Quantitative, statistical analysis of the results.

The findings from step 7 are presented in a series of research articles on each case study, supplemented by a concluding 'validation study'. The next section introduces these empirical studies.

1.4 Introduction of the studies in this thesis

1.4.1 Criteria used to select the cases to be studied

Five separate studies were undertaken, four of which are case studies while the fifth, and concluding study validates selected results from those case studies. The general criteria for a subject of a case study to be included in this study are (1) the discovery is characterised by the scientific community as a 'major step forward', i.e., a generally recognized 'breakthrough'; (2) information about the seminal publication and publications citing it should be covered by the available data sources; (3) the discoveries have led to technology as evidenced by the occurrence of patent publications that can be linked with the discovery; (4) the time lag between the publication of the seminal publication and the moment of analysis should be large enough to allow for patents publications to be available; (5) the case studies must be examples of 'charge' and 'challenge' discoveries[34]; and (6) the number of case studies is restricted to four. In Section 1.3.1 it is argued that a universal character in discoveries and inventions exists. To uncover this universal character a small number of case studies should be sufficient. Four case studies are considered to uncover enough characteristic patterns to answer the research question put forward in this study.

The four selected case studies belong to two different 'Koshland types'. One group contains 'Charge' breakthroughs and the second group 'Challenge' breakthroughs. The study on Integrase Inhibitors (Section 1.4.2) – a HIV/AIDS medicine – and the study on Graphene (Section 1.4.3) make up this first group.

[32] In the WoS 251 different subject categories describing different fields in science are defined.
[33] 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.
[34] Discoveries of the 'chance' type are not taken into account, as they are very rare as explained in Section 1.2.6

Both breakthroughs are examples of 'charge' breakthroughs. The case studies in the second group focus on two 'Challenge' breakthrough discoveries: (1) the discovery of 'Ubiquitin-mediated proteolitic system' (Section 1.4.4), and (2) the discovery of 'Introns' (Section 1.4.5). For these well-documented discoveries practically all relevant-information, in retrospect, is available. Discoveries of type 'Chance' are so rare, and supposed to be unpredictable that they are not included in this study. The study to validate the algorithms that are constructed on the basis of the results of the analyses of the four case studies is introduced in Section 1.4.6.

1.4.2 Case study #1: Integrase Inhibitors (Chapter 2 of this thesis)

In medicine, it often takes 10 to 14 years from 'upstream' disease characterization to 'downstream' drug approval for human use. In the 1980's a novel method for drug design, known as 'rational drug design' (RDD), replaced the then existing trial-and-error harvesting methods for finding new lead compounds that could act as a basis for a new medicine. The RDD success encouraged the development of various computational methods that use structural information for suggesting novel molecular structures, and this method has led to a drastic reduction in the development time of a new drug. Using the RDD route, the speed of the development of medicines increased tremendously. In this case study the development of 'Integrase Inhibitors', a type of HIV/AIDS medicine, is analysed. This study, which is presented in Chapter 2, focuses in particular on links between science and technology embodied in scientists who also act as inventors.

1.4.3 Case study #2: Graphene (Chapter 3 of this thesis)

The Nobel Prize for Physics 2010 was awarded to André Geim and Konstantin Novoselov for "succeeding in producing, isolating, identifying and characterizing graphene" (Nobel Prize Committee, 2010). Graphene seems to be an appropriate case because the publication in which the discovery was published Novoselov et al. (2004) acted as a bridge between basic scientific research and applied science, and thus marks the beginning of technological applications. Prior to this seminal publication graphene was only studied theoretically and experimentally as an integral part of graphite crystals that, as a whole, do not have the physical properties of freestanding Graphene. We discuss this case in Chapter 3, focusing on the inflow of researchers and research institutes into the field of Graphene research.

1.4.4 Case study #3: Ubiquitin (Chapter 4 of this thesis)

The original discovery of the 'Ubiquitin-mediated proteolitic system' was published in (Ciechanover et al., 1978) and was the result of the research efforts of a small team of scientists. For some years the discovery remained unnoticed

(Ciechanover, 2006) by researchers outside the small research group. The Nobel Prize in Chemistry was awarded for this discovery in 2004. This discovery is an example of a 'Challenge' type discovery and led to many R&D activities. The Ubiquitin case is an example of a discovery that in the beginning was carried by a small research team and is discussed in Chapter 4.

1.4.5 Case study #4: Introns (Chapter 5 of this thesis)

The scientific discoveries underpinning this research were published in 1977 and revealed that chromosomes comprise of a mosaic of 'exons', used to form new copies of genes, and of 'introns' that appeared to be non-functional 'interspaces'. These discoveries proved the assumption 'chromosomes for prokaryotic and eukaryotic organisms are similar in structure' to be wrong. Therefore this discovery for which in 1993 the Nobel Prize in Physiology or Medicine was awarded is an example of a 'Challenge' discovery. This discovery is an example of a discovery where after the primary discovery a second breakthrough moment occurred at a later stage, when application-oriented science took off. This case study is covered in Chapter 5.

1.4.6 Validation study (Chapter 6 of this thesis)

The results of the four case studies 'Integrase Inhibitors', 'Graphene', 'Ubiquitin', and 'Introns' were used to develop and construct five algorithms. The purpose of this validation study is to validate the algorithms, which is done by comparing the outcomes of applying the algorithms to publications published in the period 1990–1994. For the validation several external datasets such as Nobel Prizes awarded, Nature's list with the top-100 publications most cited ever, and references from review publications, patents, and social media messages were used. Publications are classified either as 'non-breakout' when it is not selected by one of the algorithms, or as a 'breakout' when it is. Based on the characteristics of the breakout publications this group is subdivided into the next four subgroups (1) 'normal' breakouts, (2) Breakthrough by proxy', (3) Science-oriented breakthrough by proxy and (4) Technology-oriented breakthrough by proxy. Further details of this classification are provided in Section 6.A.9.

1.5 Information sources

1.5.1 Research publications

The bibliographic information in the WoS contains, for every publication, the list of authors, the title, publication year, the journal in which the publication was published, the referenced publications and the publications that

cite a publication and the like. At CWTS[35] the WoS database is enriched with information produced by CWTS's internal software. A comparable data-infrastructure for the SCOPUS-database with the same quality level is not available at CWTS at this moment; therefore the in house version of WoS-database at CWTS (CWTS-WoS) is used in this study. This CWTS-WoS database contains bibliographic information of more than 44,000,000 unique scholarly publications, published since 1980. The publications have different types, and all document types are combined into four classes (1) scholarly articles, (2) letters, (3) review articles, and (4) 'other' publications. In the CWTS-WoS database, the following information, used in this study, is available for the individual publications:

- Authors — names, addresses
- Author affiliations — names, addresses
- Science field, indicated by WoS subject categories, and also by CWTS-document clusters
- Research focus of journals[36]
- Citation links between publications
- Publication date
- Title
- Abstract

This study focuses on the development and diffusion of scientific knowledge and, in particular, on the identification of scholarly publications having a more than average impact on the further development of a scientific field. Original research is published in publications that are classified in the WoS as document types 'article' or 'letter'. Review publications constitute another significant category of scientific publications and provide an overview of the most relevant research in a specific field of science. Therefore these publications do not contain results of original scientific research. Of the publications from the period 1990–1994, 76%[37] are scholarly publications, of which 91.5% are of type 'article', 6.0% of type 'letter', and 2.5% of type 'review publication'.

1.5.2 Patent publications

Patent data is used in the validation study (Chapter 6). The European Patent Office provides a database containing patent related information. This database is generally referred to as the 'PATSTAT' database[38], and is applied in the validation study (see Chapter 6). Twice a year a snapshot of the EPO master

[35] CWTS – **C**entrum voor **W**etenschap- en **T**echnologie **S**tudies (Centre for Science and Technology Studies), Leiden University, Leiden, the Netherlands

[36] At CWTS a classification is assigned to journals to express the scientific orientation of the research published in the journal. Publications inherit the science orientation of the journal. The classification consists of the six classes 'Discovery science', 'Industrially relevant science',' Science-based technological development', 'Clinically relevant science', 'Science-based clinical practice', and 'Industrial/medical development'. The methodology is explained in (Tijssen, 2010).

[37] The 24% of 'other' publications are news items, corrections, editor notes, and such like.

[38] The official name for PATSTAT is 'the EPO Worldwide Patent Statistical Database'

documentation database is taken to produce a new version of the PATSTAT database. The most recent version (Fall 2016) contains bibliographic data for over 90 million patent applications from more than 80 regional or national patent offices. For all patent publications the database contains information on:

- Patent applicants — names, addresses
- Inventors — names, addresses
- Various dates — date of application, of publication, of grant, and of lapse
- Area of technology — based on the fine grained patent classification systems
- Citation links to other patents
- Citation links to so called *non-patent literature*, e.g. scholarly publications
- The origin of a citation link — applicant, employee of a patent office, third party
- Title and abstract of an invention
- Information on patent families with equivalent documents, or families of technology related inventions

1.5.3 Measured quantities

Counting publications, citations, and patents

We use numbers of publications and numbers of citations at certain moments in time as data points. To study the evolution of science fields, the flow of publications is 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.

In the case of patents the focus is on a group of identical patent publications that describe the same invention, so-called patent families. In this study patent families that follow the DOCDB[39] definition are used to circumvent the problem of multiple counting of the same invention.

Publications of type 'article' and 'letter'

Original research is published in papers that are classified in the WoS as 'article' or 'letter'. Of these publications from the period 1990–1994 76%[40] are scholarly publications, as already mentioned in Section 1.5.1 on page 19 of which 91.5% is of type 'article', 6.0% of type 'letter', and 2.5% of type 'review publication'. As the number of review papers is small in relation to the total number of publications, not including the review papers in the analysis has

[39] Related patent publications can be grouped together in several ways. A group containing, in patent-legal sense, equivalent patent documents is called a DOCDB patent family. Patent documents in the same DOCDB-family describe the same invention.

[40] The 24% of 'other' publications are news items, corrections, editor notes, and the like.

no significant influence on the results. In this study the classification of publications in the WoS is not questioned.

Linking patents and scholarly publications

In the PATSTAT database 15% of the total number of citations in patent publications are citations to non-patent literature[41]; half of these citations are references to scholarly publications. Especially in the case of technologies that are considered very promising, like Graphene, there is often at least one publication from the USPTO present in the patent family, a group of identical patent publications that describe the same invention[42]. In this study, patent families that follow the DOCDB[43] definition are used to circumvent the problem of multiple counting of the same invention. Of all DOCDB families in the PATSTAT database, approximately 20% contain at least one USPTO patent publication. In Figure 1.1 an example of the first page of a USPTO publication (US 9,315,437) is shown. The use of patent families solves, at least in part, the problems with references to scholarly literature in patent publications (Jaffe and de Rassenfosse, 2016).

WoS subject categories and CWTS document-clusters

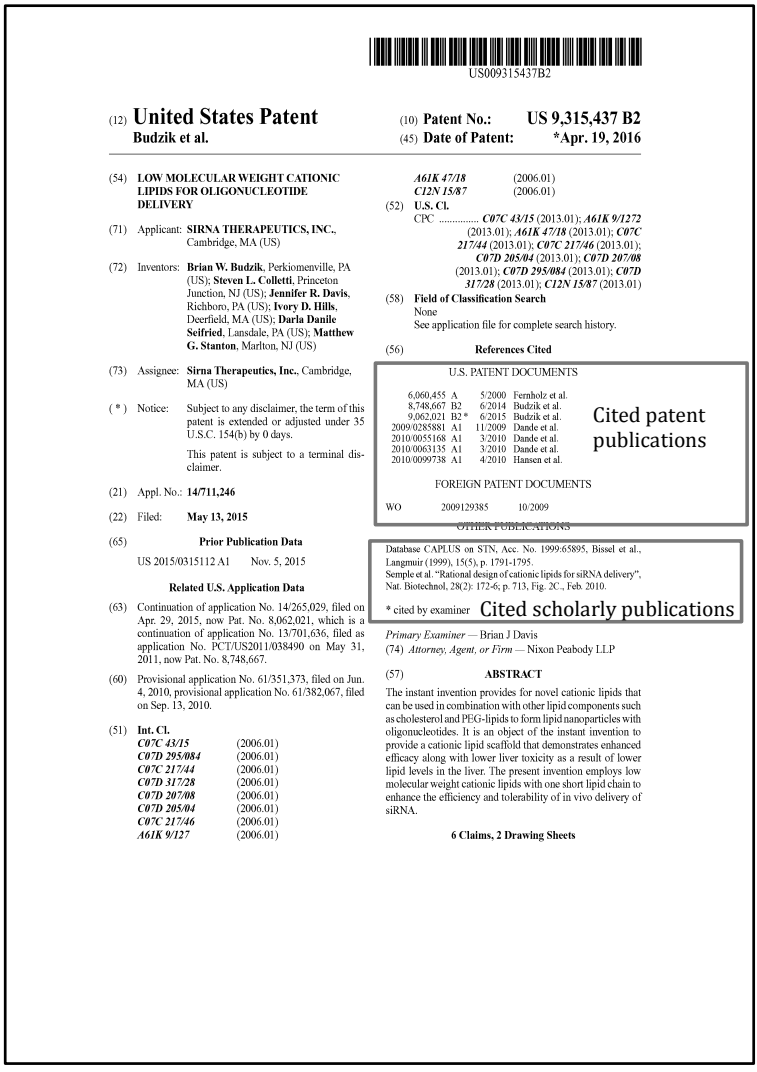
WoS subject categories group together publications in 251 different science research fields. Each of these subject categories identifies a research field in science. If the subject categories change as a result of changes in science, the bibliographic information for the publications is adapted to show the actual situation.

A document classification system has been developed recently at CWTS (Waltman and van Eck, 2012) as an alternative to the WoS Subject category system. This CWTS clustering-method groups publications together on the basis of citation relations. This CWTS-system is a hierarchical system with three levels of aggregation: micro, meso, and macro. At the meso level, the level used in this study, there are 865 different 'document clusters'. Whereas the WoS-subject categories are stable over time, the CWTS document clustering reflects the dynamic nature of science as this system is based on citation relations.

[41] Non-patent literature consists of scholarly publications, handbooks, international standards, and everything else that is considered from a patent procedural perspective to be a reference to non-patent information. Current research at CWTS shows that roughly 50% of the references to non-patent literature are references to scholarly publications

[42] Due to the US legislation patent publications published by the US Patent and Trademark Office (USPTO) contain extensive lists of references to all relevant – patents and non-patent-information. In patent publications provided by other patent offices usually only citations occurring in 'search' or 'examination' reports are mentioned; the citations mentioned in these reports constitute only a small portion of the citations in the patent documents.

[43] In several ways related patent publications can be grouped together. A group containing, in patent-legal sense, equivalent patent documents is called a DOCDB patent family. Patent documents in the same DOCDB-family describe the same invention.



US009315437B2

(12) **United States Patent**
Budzik et al.

(10) **Patent No.: US 9,315,437 B2**
(45) **Date of Patent: *Apr. 19, 2016**

(54) **LOW MOLECULAR WEIGHT CATIONIC LIPIDS FOR OLIGONUCLEOTIDE DELIVERY**

(71) Applicant: **SIRNA THERAPEUTICS, INC.**, Cambridge, MA (US)

(72) Inventors: **Brian W. Budzik**, Perkiomenville, PA (US); **Steven L. Colletti**, Princeton Junction, NJ (US); **Jennifer R. Davis**, Richboro, PA (US); **Ivory D. Hills**, Deerfield, MA (US); **Darla Danile Seifried**, Lansdale, PA (US); **Matthew G. Stanton**, Marlton, NJ (US)

(73) Assignee: **Sirna Therapeutics, Inc.**, Cambridge, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **14/711,246**

(22) Filed: **May 13, 2015**

(65) **Prior Publication Data**

US 2015/0315112 A1 Nov. 5, 2015

Related U.S. Application Data

(63) Continuation of application No. 14/265,029, filed on Apr. 29, 2015, now Pat. No. 8,062,021, which is a continuation of application No. 13/701,636, filed as application No. PCT/US2011/038490 on May 31, 2011, now Pat. No. 8,748,667.

(60) Provisional application No. 61/351,373, filed on Jun. 4, 2010, provisional application No. 61/382,067, filed on Sep. 13, 2010.

(51) **Int. Cl.**

C07C 43/15	(2006.01)
C07D 295/084	(2006.01)
C07C 217/44	(2006.01)
C07D 317/28	(2006.01)
C07D 207/08	(2006.01)
C07D 205/04	(2006.01)
C07C 217/46	(2006.01)
A61K 9/127	(2006.01)
A61K 47/18	(2006.01)
C12N 15/87	(2006.01)

(52) **U.S. Cl.**

CPC ***C07C 43/15*** (2013.01); ***A61K 9/1272*** (2013.01); ***A61K 47/18*** (2013.01); ***C07C 217/44*** (2013.01); ***C07C 217/46*** (2013.01); ***C07D 205/04*** (2013.01); ***C07D 207/08*** (2013.01); ***C07D 295/084*** (2013.01); ***C07D 317/28*** (2013.01); ***C12N 15/87*** (2013.01)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,060,455	A	5/2000	Fernholz et al.
8,748,667	B2	6/2014	Budzik et al.
9,062,021	B2 *	6/2015	Budzik et al.
2009/0285881	A1	11/2009	Dande et al.
2010/0055168	A1	3/2010	Dande et al.
2010/0063135	A1	3/2010	Dande et al.
2010/0099738	A1	4/2010	Hansen et al.

Cited patent publications

FOREIGN PATENT DOCUMENTS

WO	2009129385	10/2009

OTHER PUBLICATIONS

Database CAPLUS on STN, Acc. No. 1999:65895, Bissel et al., Langmuir (1999), 15(5), p. 1791-1795.
Semple et al. "Rational design of cationic lipids for siRNA delivery", Nat. Biotechnol, 28(2): 172-6; p. 713, Fig. 2C., Feb. 2010.

* cited by examiner

Cited scholarly publications

Primary Examiner — Brian J Davis
(74) *Attorney, Agent, or Firm* — Nixon Peabody LLP

(57) **ABSTRACT**

The instant invention provides for novel cationic lipids that can be used in combination with other lipid components such as cholesterol and PEG-lipids to form lipid nanoparticles with oligonucleotides. It is an object of the instant invention to provide a cationic lipid scaffold that demonstrates enhanced efficacy along with lower liver toxicity as a result of lower lipid levels in the liver. The present invention employs low molecular weight cationic lipids with one short lipid chain to enhance the efficiency and tolerability of in vivo delivery of siRNA.

6 Claims, 2 Drawing Sheets

Figure 1.1: Front page of United States patent 9,315,437