Moving Beyond Bibliometrics:
Understanding Breakthrough Emergence through Missed Opportunities

Sen Chai
Harvard University
chais@nber.org

October 7, 2013

DRAFT – Please do not cite without author permission

ABSTRACT

Who is most likely to discover a breakthrough? Why are some scientists more successful than others at discovering them? By using extant theories of breakthrough emergence to predict a groundbreaking discovery in biology, RNA interference, I show that the explanatory power of combining all current theories are weak because they sample on rare successes rather than the multiple instances of failure in the discovery process. Instead, I focus on understanding these failures by interviewing scientists with high potential of discovering breakthroughs in a case historical analysis. My findings suggest that the seminal discovery was missed several times not only due to difficulties in solving a particular problem but also due to failures in identifying breakthrough opportunities. I propose a cognitive framework with institutional underpinnings at the basis of these failures. In the problem identification stage, framing barriers from pursuing normal science and existing boundary barriers between communities of scientists contribute to difficulties in identifying the breakthrough opportunity by misrepresenting the magnitude of the problem. In the problem-solving stage, scientists are constrained by paradigmatic pressures to avoid being wrong, and coupled with boundary barriers similar anti-dogmatic observations stay isolated and unsubstantiated, thus diminishing confidence to identify a new revolutionary paradigm.

Keywords: breakthrough emergence; missed breakthrough; bibliometric measures; hybrid methodology

Acknowledgements: I would like to thank the Harvard Business School Department of Research and Doctoral Program for supporting my work. I am also grateful to Lee Fleming, Richard Freeman, Luciana Silvestri, Ethan Bernstein, Hila Lifshitz, Tiona Zuzul, Gary Pisano, Fiona Murray, Vicki Sato, and members of the HBS CRAFT course, for their insightful comments and feedback.

INTRODUCTION

Starting with Schumpeter's notion of creative destruction (1942), scientific and technological breakthroughs have intrigued scholars and practitioners alike. The literature is rife with works that attempt to identify sources of such breakthroughs, especially when radical discoveries and inventions prove to be an important foundation of scientific and technological advancement, and are linked to wealth creation and economic growth. Even though being able to forecast who is most likely to discover a breakthrough[1] is important for managers in helping them identify key scientists and to inform policy on public investment in science, combining extant theories of breakthrough emergence to predict a groundbreaking discovery in biology, RNA interference, yielded weak explanatory power.

The lack of predictive power hint that current literature's study of what factors foster breakthrough emergence by sampling on rare successes from archival sources only contributes to a partial understanding of the phenomenon. This stems from breakthroughs being inherently rare and serendipitous events, and involving cognitive mechanisms difficult to capture using pure bibliometric archives. Thus, there is much to be gained in digging deeper into the counterfactual process of why and how breakthrough discoveries are often missed and delayed. However, such failures cannot be easily observed and measured with current methods, thereby limiting the ability to make inferences and forcing the literature to focus on successes. The bias in emphasis on the final successful outcome of breakthroughs emerging also drove the literature to concentrate most of its efforts on finding what enhances the end goal of solving a particular

[1] Breakthrough is defined herein following Simonton's notion of impact (1999) encompassing dimensions of both creative novelty and success. As opposed to some discoveries or inventions that become scientific or technological dead ends, breakthroughs are advances that disturb the previous understanding of a particular phenomenon in a fundamental manner and are foundationally at the basis of further enhancements.

problem without taking into account that failures also exist throughout the process, such as at the problem identification stage. Thus, this paper departs from prior works in three novel ways: first, it digs deeper into the counterfactual process of why and how breakthroughs are missed and delayed; second, it uses a case historical analysis of RNA interference with qualitative interview data thereby offering a glimpse into the informant's train of thought and sense making and bypassing the issue of unobservable failures with archival data; and third, it investigates failures throughout the discovery process not only focusing on the end goal. The paper answers the research questions of who misses breakthroughs and why and, in turn, sheds light on why some scientists are more successful than others at discovering them by taking this counterfactual lens.

I find that scientists on the verge of breakthrough missed the seminal discovery not only due to difficulties in solving a particular problem, but also because of failures to identify the breakthrough opportunity. At the basis of this failure underlies a cognitive mechanism stemming from three barriers embedded in institutional norms between science and technology: (1) framing barriers – impediments that prevent scientists from considering a phenomenon under study from different viewpoints, (2) boundary barriers – obstacles that hinder the transfer of knowledge between various scientific fields, and (3) paradigmatic pressures – forces that constrain scientists from proposing drastically different or even contradictory theories. In the problem identification stage, paths dependence from established technologies and the quest toward normal science blinded scientists from recognizing a breakthrough potential. Instead they framed RNAi as a tool useful in uncovering answers to their initial experiments and ignored it as a scientific concept worthy of study in and of itself. Furthermore, boundary barriers between communities of scientists hindered repeated instances of odd observations in prior works to be connected together, thereby misrepresenting the magnitude of the problem and impeding identification of

the breakthrough opportunity. In the problem-solving stage, scientists were constrained by the confines of current dogma. Pressured by established paradigms, they hesitated to propose solutions that significantly strayed away from existing theory to avoid being wrong. Again coupled with boundary barriers that prevented linkages, similar anti-dogmatic observations and results stayed isolated and diminished scientists' confidence in identifying and proposing a new revolutionary paradigm.

The organization of this work is as follows: I review the literature on sources and processes of breakthrough emergence, and synthesize them by testing the predictive power of these current theories. Then, I place the RNAi discovery in historical and scientific contexts and describe the methods I employ to compile a dataset on the discovery of RNA interference. In the paper's core, I propose a cognitive framework that explains why breakthroughs are missed at various stages of the discovery process. Finally, I conclude with a discussion on the implications of my results to extant literature.

LITERATURE REVIEW

Scholars of innovation have put forth many hypotheses identifying sources of creativity starting, at the organizational level, with the debate between whether small entrepreneurial entrants (Schumpeter, 1934) or major incumbents (Schumpeter, 1942) are the basis of creative inventions. It expands into identifying capabilities that are required of firms to stay inventive, such as absorptive capacity (Cohen & Levinthal, 1990), dynamic capabilities (Teece, Pisano, & Shuen, 1997), experimenting early and often (Thomke, 2003), and sampling a large landscape for multiple trials (Rivkin & Siggelkow, 2002) and recombination (Fleming, 2001). These works have mainly concentrated on the firm's ability to build problem-solving skills as a way to sustain

creativity, and have largely ignored the significance of problem identification in creating breakthroughs that I stress herein. Moreover, these studies cannot be readily applied to the current context because the locus of decision-making in scientific research is centered at the principal investigator level where heads of labs are responsible for providing funding, hiring personnel, deciding research direction, etc. This predominant structure in science, thus, prevents analysis at the organizational level and requires studies to be performed either at the team/laboratory or individual level. At the laboratory level, students or fellows initially working in the same lab eventually take on professorship positions. The dynamic nature of the boundaries of these groups precludes analysis at this level; therefore, this work centers on the individual scientist.

At the individual level, the literature has mainly focused on identifying (sometimes conflicting) factors or characteristics that enhance breakthrough discovery. Failure to reach consensus can be attributed to lack of experimental evidence and reliance on correlational studies. Despite recent and laudable emphasis upon causality, too few conclusive studies have been yet published. For instance, being more productive increases the number of creative draws thereby improving chances for breakthrough discovery (Simonton, 1999), whereas being less productive may also improve those chances by focusing and pursuing anomalies. Social brokers – those who are the sole connections between others – have shown to be more creative (Burt, 2004) benefitting from first access to and control of information, less creative (Obstfeld, 2005; Uzzi, 1997), and more creative in particular circumstances but also hampered in their ability to diffuse their idea (Fleming et al. 2007). Individuals at the core of a community are a more likely source because they enjoy enhanced information and resource access from social ties (Collins, 1998; Gieryn & Hirsh, 1983); or at the periphery because they are not constrained by prevailing

assumptions and theories (Jeppesen & Lakhani, 2010). Specialists with deep technical knowledge are better equipped to see beyond the frontier and make more accurate predictions, as opposed to generalists who can bring together disparate components (Dougherty, 1992; Leonard-Barton & Swap, 1999). Individuals realize breakthroughs earlier in their careers because they are not constrained by the thinking of their field (Simonton, 1989), or later because they must work through the accumulation of knowledge (Jones, 2009). Better scientists prefer to stay in academia as they value the freedom in choosing their research direction (Stern, 2004), yet some corporate labs also do fundamental breakthrough work. Affiliation with prestigious institutions increases breakthrough potential because of higher human capital and exposure to better ideas. Mobility between multiple affiliations increases exposure to a greater diversity of ideas, but is associated with high setup costs and may also be an indicator of failed tenure attempts (McEvily & Zaheer, 1999). Collaboration might increase the chances of breakthroughs because it increases the diversity of search and efficiency of idea selection (Singh & Fleming, 2010; Wuchty et al., 2007), though working individually at some points in the process also appears beneficial because it minimizes idea suppression and social loafing (Girotra et al., 2010).

Despite continued interest in the literature, no study has assessed the extent to which altogether existing sources can accurately predict a future breakthrough. Combining the scientists' bibliometric attributes – brokerage vs. cohesion, periphery vs. core, specialist vs. generalist, experience vs. youth, affiliation type, affiliation prestige, mobility, and lone scientists vs. teams – before 1998, I predicted the outcome of who would have a breakthrough year in 1998. I operationalized breakthrough using four measures from the most stringent depiction of being an author on the Nobel paper, and gradually relaxed it to being in the top ten percent of the citation distribution, to the number of forward citations received, and finally to the number of

publications. The regression models I employed for these predictions mirror the outcome measure. For instance, the rare event logistic (relogit) model was used to predict authoring the Nobel paper; the logistic model with cluster robust standard errors was used for scientists who were in the top ten percent of the citation distribution; and finally, quasi-maximum likelihood Poisson models with robust standard errors were employed when the forward citation counts of 1998 papers and the number of publications in 1998 were the outcome variables. Moreover, using OLS I assessed the predictive power of all existing sources of breakthrough combined. Appendix A contains detailed descriptions of each variable and model, as well as interpretations of the results.

The effect size from each source of breakthrough is summarized in Figure 1, and shows that prior productivity, prior eminence, brokerage, youth and newcomer consistently contribute to a researcher's subsequent impact (measured using the top tenth cites and citation count). From OLS models, the percent contribution to the variance of each explored theoretical theme in all four models is depicted in Figure 2. Altogether current bibliometric theories on creativity can explain only 0.6 percent of the variance in predicting the Nobel paper (compared to ~0.15 percent random chance that the six are authors of the Nobel winning paper), 5.3 percent for authoring a paper in the top ten percent of citations, 13.1 percent for citations, and 13.4 percent for publications. Thus, our ability to predict from whom breakthroughs are most likely to emerge is still relatively weak. And if prior publications and citations are included, these numbers increase to 0.8 percent, 20 percent, 37.8 percent, and 49.6 percent, respectively.

[Insert Figures 1 and 2 about here]

Aside from factors that foster breakthrough discovery, the process by which technological inventions and scientific discoveries emerge has also been much subject of study.

In the technological realm, the emergence of a standardized technological form from multiple paths is thought to be either socially (Bijker, Hughes, & Pinch, 1987) or socio-cognitively constructed (Garud & Rappa, 1994). The focus is on how standardization is achieved amongst multiple prospective technological forms, different from the process of breakthrough emergence where scientists strive to find a single truth. Within the scientific institution, Kuhn introduced the notion of paradigm shifts as the underlying mechanism for scientific revolutions through accumulated anomalies that mount to crisis (1962). However, the work is mainly theoretical and describes processes by which scientific revolutions emerge rather than focusing on what barriers impede them. I build onto this work by delving into the counterfactuals and provide detailed mechanisms of failure that cause delayed and missed scientific revolutions, complementing the few existing works that take this counterfactual perspective (Berson, 1992; Dyson, 1972).

RNA INTERFERENCE

RNA interference is a naturally occurring endogenous gene silencing mechanism that strays away from the central dogma of molecular biology that dictates how genetic information encoded in double-stranded DNA unzips, transcribes into RNA and finally translates into protein. It is triggered by double-stranded RNA (dsRNA) precursors (Fire et al., 1998), and can ultimately turn genes on and off through specific genetic interference mechanisms. This potent causal agent was identified from Andrew Fire and Craig Mello's breakthrough insight that preparations of single-stranded sense and antisense RNA in test tubes were contaminated and annealed into dsRNAs (Fire, 2007). For this discovery the two were awarded the Nobel Prize in Physiology and Medicine in 2006. RNA interference was also coined as a result of this work. RNAi is valuable as a research tool and in biotechnology therapeutic development. In research,

synthetic dsRNA introduced into cells can induce suppression of specific genes of interest both *in vitro* and *in vivo*, thus enabling scientists to understand gene function. In therapeutics, RNAi pathways can be conceivably used to treat genetic diseases.

The history of RNAi is a story of how several seemingly unconnected and unexpected phenomena observed in various organisms across kingdoms were finally linked together after discoveries to the trigger and underlying mechanism were made. As it turns out, RNAi is a fundamental mechanism that dates back millions of years where single-celled organisms cleverly employed it to defend themselves against the invasion of foreign viruses. Its modern day discovery started in the late 1980s and early 1990s in plants. At that time plant biologists were attempting to transgenically alter color in petunias by introducing an enzyme that encodes pigmentation in flowers. When the experiment was initially designed the expectation was to see gene overexpression manifested through darker colors (Krol, Leon, Beld, Mol, & Stuitje, 1990; Napoli, Lemieux, & Jorgensen, 1990). Instead to everyone's surprise, the petunias became less pigmented than their natural form producing fully or partially white flowers. This indicated that as opposed to the intended gene overexpression, activity of the enzyme had significantly decreased expression to the point of deactivating the gene responsible for regulating color pigmentation. However, both the underlying mechanism and trigger remained unknown.

The story then moves to the fungal community where independently a similar phenomenon was observed by scientists studying neurospora crassa fungi (Romano & Macino, 1992) and was separately named quelling. A few years later in the c. elegan worm community, scientists attempting to understand the purpose of a particular gene in embryo cells found much like co-suppression in plants that not only did the single-stranded RNA antisense silence the gene under study so did the corresponding sense RNA strand that was designed as negative control

(Guo & Kemphues, 1995). Not long after, plant virologists also found a similar unexpected phenomenon when attempting to improve plant resistance from viral infections that they labeled virus-induced gene silencing (Ratcliff, Harrison, & Baulcombe, 1997).

Although all of these odd observations were not immediately recognized as related to one another, each community, plant and animal scientists, were all independently aware of the phenomenon prior to the discovery of its trigger in 1998 by Fire and Mello. The European plant community at the beginning of the 1990s had already started their own network of laboratories with the aim of joining together and applying for funding to study the phenomenon. In the animal community, more specifically the c. elegan worm community, many had come across the phenomenon in their own experiments while being unaware of the intricacies of the underlying mechanism. Others although not always getting consistent, reproducible and potent results used the precursor technology to RNAi, antisense oligonucleotides, as a tool to inhibit and study the function of specific genes. It was also the topic of discussion at several conferences around that time such as the Pew Scholar workshop.

METHODS

Identification of a Community of Scientists

How a community is defined is crucial to understanding how scientific breakthroughs arise within it. The RNAi community is defined functionally where I include content search of titles, abstracts and Medical Subject Headings (MeSH) keywords, and used in determining the sample for both predictive regressions presented in the literature section and qualitative interviewees. Because my study is centered on the period prior to breakthrough where a defined

community of RNAi researchers had yet to emerge, keywords such as "RNA, Interference", as well as "co-suppresion" and "quelling", the same phenomenon in plants and fungi, did not enter the MeSH lexicon until 2002. To bypass this issue, I reviewed archival documents on the history of RNAi including the Nobel lectures, and found that scientists sought to explain gene expression regulation or gene silencing by experimenting with both dsRNA and antisense RNA as causal agents. Furthermore, they believed in the premise that RNA molecules are not only restricted to the passive role of carrying genetic information but also possess catalytic functions, thus leading to the hypothesis that RNA plays a central role in gene silencing mechanisms. Consequently, I defined the community of researchers with the potential of discovering a breakthrough from their published peer-reviewed articles using the MeSH search terms[2] "RNA, Double-Stranded", "RNA, Antisense", "RNA, Catalytic", "Gene Silencing" and "Gene Expression Regulation" in PubMed. I augmented the MeSH search with title and abstract searches to include scientists who initially observed the RNAi phenomenon in plants and fungi[3]. I incorporated papers that were published until 1999, as those who quickly followed the 1998 breakthrough paper were also in the risk set for breakthrough discovery.

By extracting unique authors from the set of papers obtained above, I identified a community of RNAi scientists. This sample of scientists defining the pre-breakthrough RNAi community yielded 1,551 papers and 3,959 unique authors. However, due to missing data for 49

[2] The exact search string used in PubMed query extracted on October 26, 2011: ((((gene silencing[MeSH Terms] OR gene expression regulation[MeSH Terms]) AND (RNA, double-stranded[MeSH Terms] OR rna, antisense[MeSH Terms] OR rna, catalytic[MeSH Terms])) AND "1980"[Publication Date] : "1999"[Publication Date]) AND English[Language]) NOT interferon[MeSH Terms]. I also found that dsRNA generated a lot of noise as it was heavily used by immunologists studying interferon responses. To minimize the noise from interferon I include in the MeSH search the "NOT interferon[MeSH Terms]" term.

[3] The exact search string used in augmented PubMed query extracted on October 26, 2011: ((((cosuppression[title/abstract] OR co-suppression[title/abstract] OR quelling[title/abstract] OR RNAi[title/abstract] OR RNA interference[title/abstract])) NOT interferons[MeSH Terms]) AND "1980"[Publication Date] : "1999"[Publication Date]) AND English[Language].

such individuals, the sample shrunk to 3,910 authors. Out of the 3,959 unique authors present in the sample, 144 authors do not have any prior publications either within the RNAi community or any other tangential field within the life sciences.

Data Sampling and Collection

Following prevailing assumptions in the innovation literature that highly uncertain creativity is a path dependent process of recombinant search rather than a single radical event (Fleming, 2001; Henderson & Clark, 1990), I conceptualized breakthrough as marked by multiple failures before eventual success. Therefore using a case history method to study breakthrough emergence was appropriate (Corbin & Strauss, 2008; Miles & Huberman, 1984), as it unearthed the nuances of multiple trials along the path of discovery irrespective of whether they were failures or successes. Since I studied the breakthrough *ex post*, understanding the circumstances scientists faced *ex ante* is critical. Although interviews potentially suffered from hindsight bias, they were useful in inquiring about causes of failure that were hard to obtain using purely archival methods. To minimize retrospective sense making, I triangulated my findings from the interview data with archival sources such as the Nobel lectures, transcriptions of the Nobel interviews, and RNAi paper publications from each interviewee (Golden, 1992).

Many historical case studies of breakthroughs exist. Although rich and descriptive when characterizing the invention or discovery, the number of stakeholders included in such historical accounts is usually limited to those in the immediate proximity of the winners, such as their mentors, collaborators, and eminent fellow scientists racing for the same discovery. These individual historical accounts may suffer from convenience sampling and lack the macro and systematic view enabled by large archival quantitative methods. To ensure exhaustiveness of my

case history, I developed a selection method through residual analysis to systematically determine not only the researchers who emerged but also ones with the highest potential. I concentrated my interviews on those who did not ultimately discover the breakthrough – the counterfactuals – to gain an understudied perspective on the phenomenon.

The interview process consisted of two stages. I first interviewed two individuals, a then board member of a leading RNAi technology based company and a scientist familiar and knowledgeable about RNAi and its history but not doing research in the area. These two interviews allowed formulation of questions in preparation for the main round with the actual actors involved in its discovery. The interviews in this first stage lasted an average of 30 minutes. They were semi-structured and discussions centered on how to define the community of scientists focusing on RNAi as well as the discovery trajectory. For instance, one interviewee pointed out that rapid development in the field of molecular biology and genetic engineering, such as DNA sequencing, recombinant DNA and the hypothesis that life originated from RNA (Gilbert, 1986) were precursors to the discovery of RNAi. Furthermore, they both brought my attention to the fact that most historical expositions of RNAi tended to include observations in plants and fungi. However, it was not clear whether researchers working with animal models at the time were aware of or even associated their work to these prior anomalous results found in the plant systems. These conversations fine-tuned existing interview questions and triggered new ones for the subsequent set.

My method of selection for interviewees for the second stage of interviews built on the regression model presented in the literature review of this work that predicted the citation count of 1998 publications given prior bibliometric characteristics of each scientist in the sample. I ran residual analysis on the citation count of 1998 publications model to identify scientists who were

incorrectly predicted. I calculated the error term by taking the difference between the predicted publication impact, *E(Y),* and the actual number of forward citations for 1998 publications, Y_i, as depicted graphically in **Error! Reference source not found.**. The group of interviewees consisted of the top and bottom one percent of the error terms – elite scientists who the model severely failed to predict accurately. While no interviewee was part of the Nobel winning team for RNAi, this sampling provided me with counterfactual accounts by those who had the potential to make the groundbreaking discovery but ultimately missed it.

[Insert Figure 3 about here]

I identified a total of 19 scientists with research focus in RNAi. The reason for such a low number of interviewees from the initial large sample of authors used in the regression models is two-fold. First, my technique of selection using residuals cut out a vast majority of the initial sample because I was solely pinpointing extreme cases of individuals who were predicted inaccurately. Second, because I was studying a nascent field that significantly grew in size only after the breakthrough occurred the number of scientists at the beginning was extremely limited. According to one of my respondents, the field evolved toward 500 members in the 15 years following the seminal publication.

I augmented this set of interviewees with those who attended RNAi related conferences – Keystone Symposia and Gordon Conferences – at the beginning of the field, which increased the number of interviewees to 27. Of these 27 scientists I reached out to with interview requests 18 responded positively. During the interviews, I inquired about other potential individuals within the RNAi community at its birth stage that my interviewee would recommend I meet to validate my selection technique. Their suggestions were all amongst the sample of 27 interviewees I identified with the selection methods described above. Respondents spanned model organisms

from plants, worms, fruit flies, to humans and included geneticists, molecular biologists and biochemists that contributed to RNAi research conceptually and technologically. They also included one Nobel Prize winner and three Lasker award winners. Each interview lasted between 60 to 120 minutes, averaging 75 minutes.

Interview questions were semi-structured such that open-ended questions were asked first, followed by more specific and probing ones. I started by inquiring about the line of research each informant was undertaking during the period shortly before the 1998 breakthrough was made and covered several other topics from understanding circumstances around and factors leading to breakthrough discovery, to defining and characterizing the community of scientists prior to emergence of the RNAi field. The interview guide is shown in Appendix B.

Data Analysis

Analysis of each interview once transcribed verbatim was conducted in line with coding principles set out by qualitative researchers (Miles & Huberman, 1984). I first open coded all interviews by describing each excerpt, such as 'attended conference', 'ignored mechanism', 'used RNAi as tool', 'double-checked in another organism', 'described antecedent to RNAi', 'explored at the fringe', etc. When new data did not fit a previously identified code category I created a new category. Once I finalized the open code for all primary interviews, I proceeded to axial code the open code categories. Two salient axial code classes emerged: barriers to breakthrough and actions scientists took to circumvent barriers. A third category included all other breakthrough related narratives such as the historical context. The two salient classes were further divided into three barriers (as well as instances where the barriers interacted with each other) that correspond to the three themes that finally emerged: being blinded by conventional

science from framing barriers, being constrained by current dogma from paradigmatic pressures, and being unable to connect the dots due to boundary barriers. I also obtained for each theme a collection of practices that scientists put in place to circumvent barriers to breakthrough.

FINDINGS

The salient themes that emerged highlight how a number of scientists were on the verge of breakthrough several times but missed the seminal discovery. In other words, these results centered on uncovering explanations behind the counterfactual of missing breakthroughs. My novel finding is that delay in discovery was not only due to struggles in solving a particular problem but to difficulties in identifying the problem, in assessing the potential impact of the problem as well as in proposing a drastically different theory than stipulated by the current paradigm. Thus, throughout the discovery process that I divided into problem identification and problem-solving stages, those on the verge of discovery suffered from failures to identify and propose breakthrough opportunities. The results suggest that at the basis of this failure underlies a cognitive mechanism hinged on institutional underpinnings stemming from three barriers – framing barriers, boundary barriers, and paradigmatic pressures – and their interactions. Framing barriers are impediments that keep scientists from considering a phenomenon with different viewpoints. Boundary barriers are obstacles that hinder the transfer of knowledge between various scientific fields. And paradigmatic pressures are forces that constrain scientists from proposing new drastically different or even contradictory theories. These findings also shed light on how the scientific and technological institutional logics differ, and how the divergent nature of knowledge produced between the two is manifested in the discovery of scientific breakthroughs. During problem identification, scientists who missed the opportunity

pursued normal science by framing anomalous observations along established technological paths. During problem solving, they were held back by paradigmatic pressures and interpreted abnormal results according to established paradigms. Existing boundary barriers between communities of scientists compounded both effects as they prevented similar anomalous patterns from being connected together, thus hindering pattern recognition and pattern labeling. Figure 4 illustrates this framework graphically, while Table 1 contains quotations illustrating each barrier from all 18 respondents that I interviewed.

[Insert Table 1 about here]

[Insert Figure 4 about here]

Problem Identification Failures

Framing Barriers

Because most scientists came in contact with the phenomenon of RNAi as a technique to silence genes in pursuit of hypothesis driven science, their views of the phenomenon were biased toward using it as a tool rather than a topic of inquiry worthy of scientific merit. Path dependence from prior technologies reinforced the belief that the phenomenon of gene silencing was a technique, which cognitively biased and ultimately delayed discovery of its trigger. Underlying institutional logics in science triggered this cognitive bias because researchers were blinded by the pursuit of normal science in their quest for more publications. Being able to use the technique to accomplish the end goal of inhibiting specific genes mattered more than understanding why the technique worked. Thus, researchers were unable to identify the interesting and potentially groundbreaking problem to be solved, and subsequently passed on the breakthrough opportunity. As described by a respondent below,

> "My sense from [others] was that they just looked at this like a bizarre tool, they couldn't explain it but it was fabulous for what they wanted to do. They could silence genes. […] They were focused on the thing at hand and kind of ignoring this elephant in the room, which was far more important and interesting." (respondent 6)

Most researchers valued the phenomenon's ability to inhibit specific genes without having to rely on mutations. It was a means to an end rather than the end itself. This behavior is in line with Kuhn's (1962) prediction that scientific research is extremely productive at expanding the central paradigm but also self-reinforcing during periods of normal science. Case in point, the two scientists, Guo and Kemphues, who first observed the phenomenon in 1995 in worms explicitly chose not to study why it worked. One of the two scientists explained following their observation of the anomalous gene silencing phenomenon, "once we knew it was a gene specific effect we didn't really care how it worked. All we cared about was that we could use it." (respondent 10) They reported the strangeness that the control in the experimental design, sense RNA, had a similar potent effect as the treatment, antisense RNA, in silencing a gene they were studying, but decided that it was not worth following up.

Contrary to technological innovations where a breakthrough invention happens at first successful occurrence, a number of scientists were on the verge of breakthrough. Novelty in science is not mere observation but lies in the explanation of why a particular result occurs. In other words, the definition of success in science is different from that in technology. Hence what would have been a success in the technological realm is not considered as one in science because observations or descriptions of a phenomenon are insufficient.

This cognitive bias of seeing the phenomenon as a tool rather than conceptually understanding how it worked stemmed from the historical context of precursor technologies. In the late 1980s, large groups cornered the market in being able to produce knockout mice and

made it prohibitively hard and expensive for small laboratories to obtain such knockout samples for research purposes. Therefore, a large community of researchers was hoping for antisense oligonucleotide technology to be the key because it meant they could do things much faster than by mutation.

This barrier also explains why surprisingly little racing was present in the community to solve the puzzling mechanism. Because RNAi was perceived as a technique of how rather than a demonstration of why, scientists did not consider solving the intrigue around the RNAi phenomenon as a priority-based incentive along the institutional norms of science (Merton, 1957). They were preoccupied with other scientific endeavors that met this criterion more explicitly. Besides Fire and Mello's groups working with c. elegan worms and actively attempting to solve this puzzling gene silencing phenomenon, only plant scientists were trying to explain the same mechanism (Waterhouse, Graham, & Wang, 1998). Competition, instead, intensified *after* the pathway's trigger was found, as described by the two respondents – the first working on animal models and the second working on plant models – below,

> "For the actual initial discovery that you can introduce duplex RNA into cells to specifically inactivate genes, Fire and Mello were ahead of the game in that case. But once that discovery was made and the transition made to studying the mechanism and the factors involved, that's when the real competition came in." (respondent 5)

> "At the end of the 90s and beginning of 2000 it was really difficult, because all the things that could be found simply were found at the same time, in a range of a few months." (respondent 14)

Following Fire and Mello's discovery of its trigger, RNAi was now established to be an open and interesting scientific question to research and was now in line with the norms of the scientific institution (Dasgupta & David, 1994; Merton, 1957), as assessed by a respondent,

> "What Fire and Mello did is that they discovered that RNAi was real biology. Because, first of all, most people thought that the silencing phenomenon back then reported in

> plants and in worms, were weird things that would probably turn out to be artifacts later and they have the feeling of homeopathy." (respondent 16)

When unexpected results appear in tangential elements not affecting core hypotheses of the research project, whether manifested in the tool or the experimental results, the decision of whether to follow and inquire deeper into a weird but interesting observation or to stay with the experiment at hand is difficult. In particular, time and resource constraints together with the low probability that the oddity will eventually turn out to be something influential make it an especially hard decision, as often times they turn out to be mere artifacts. Consequently, in pursuit of normal science most ignored weird observations and carried on. However, whenever such abnormal observations occur, it is often precisely under these circumstances where breakthroughs are most likely to be discovered. As the Nobel laureate I interviewed described,

> "When you have a well-defined system and it's telling you something you don't understand, it isn't consistent with the way you've designed the system then something is new in the system. It's paying attention to that [bizarre phenomenon] and not pushing it out of the way as you went towards your more conventional hypothesis driven science. There is a new science there. To ignore that, to do conventional science is what most people will do. […] That meant the difference between the genius and good science" (respondent 13)

Boundary Barriers

Also present in the problem identification stage is the boundary barrier between disparate scientific communities. The history of RNAi's discovery is punctuated by several documented observations of the bizarre phenomenon first in plants (Napoli et al., 1990), then in fungi (Romano & Macino, 1992), worms (Guo & Kemphues, 1995) and plant viruses (Ratcliff et al., 1997), and perhaps even more instances of undocumented observations before the underlying trigger agent was finally found. Tracing through citations that the latter three papers refer to, I

found a clear dichotomy between the plant/fungus scientists and the worm scientists, where scientists from the two communities only cited within their community but not across.

The citation patterns and the independent results stemming from the plant and animal communities suggest that information flowed easily within each community not between communities. Although several observations of a similar anomaly were made in various organisms and fields, they were brushed away as a weird phenomenon that happened in the particular model organism. Thus, these boundary barriers led to two levels of discontinuity, either scientists did not know of prior anomalies from different communities or they did know but were unable to make the connection between them. The following quote by an animal scientist highlighted the latter discontinuity.

> "[Two plant scientists] were telling me stories about silencing they put in. They had these flower color things trying to get purple it would turn white, it was all screwed up. But I missed it entirely. I did not see the connection." (respondent 12)

As a consequence, scientists were unable to connect the dots and identify a repeated pattern of weird results that would provoke crisis and revolutionary changes to the established scientific paradigm (Kuhn, 1962). Had links been made between similar observations in different organisms the likelihood of dismissing an observation as odd would have been lower.

Just like cell membranes in biology, organizational boundaries form natural barriers to the diffusion of information (Kogut & Zander, 1992). For the open community of science, the flow of knowledge is deterred by boundaries of various scientific communities, which in turn hindered scientists' ability to connect the dots. For instance, aside from citation evidence presented above, most informants who worked with animal model organisms were unaware of the research done by plant and fungal scientists, and vice versa. Prior to Fire and Mello's discovery in 1998, collaboration and communication between plant and animals scientists

working on gene expression and inhibition were little to none. Most researchers from the disparate model organism communities did not meet until after the links between their works became obvious. Two animal scientists described how they perceived the plant community.

> "The plant world tends to have its own group of people, and they don't tend to intermix too much with the non-plant people." (respondent 11)

> "I am on the virus division for the society of microbiology so our job is to organize annual meetings in virology for Europe, and we have a plant virus section. But they might as well be lectured in Latin; they don't really integrate with other people. […] RNAi is an absolute classic, they had stuff going on they probably thought it was very interesting but they probably didn't think others would want to know. They didn't think animals did the same thing." (respondent 7)

Analogously from the perspective of a plant scientist, the divide between the two communities was described in a similar fashion. In fact, the plant community working on solving gene silencing was surprised that Fire and Mello's paper was published in Nature, illustrating how disparate the two communities studying the same gene silencing phenomenon were prior to the 1998 RNAi breakthrough.

> "But for the animal people, I understand that it was really a breakthrough to consider that long dsRNA could trigger something because they were not anticipating this. In plants it was not really a breakthrough, it was completely expected. That's why the discovery in the same year, in 1998, by two groups that dsRNA could induce very efficiently silencing and actually much more efficiently than sense and antisense directly was just the next step of something that was going on for 10 years. […] In plants it was more continuous." (respondent 14)

Interplay between Framing and Boundary Barriers

During problem identification, not only were both barriers present but the interaction between the two also exaggerated the effect of misidentifying the problem. Prior established antisense technologies, which was used as a tool to knockout genes, influenced most scientists to frame the phenomenon as a technique and led them to ignore it as an interesting subject of

scientific inquiry. This failure in detecting the right problem to study was further compounded by boundary barriers between scientific communities that prevented anomalous observations from various fields to be linked together. Scientists were unable to recognize a repeated pattern of anomalies. The lack of critical mass and skepticism in one single anomalous instance contributed to the inaccurate assessment of the impact of the potential breakthrough.

The top portion of Figure 4 summarizes pattern recognition failure in the problem identification stage. As illustrated by one respondent from the plant community describing an encounter where she discussed the gene silencing mechanism with an animal scientist in the mid-1990s,

> "I remember talking to a guy in Vienna who was one of the first big people making transgenic mice and he just scoffed at the whole idea that you'd see something like that. But even at the time […] there still wasn't a lot of people coming together and thinking that might all lead to a single mechanism." (respondent 18)

In fact, most of the animal community up until the Fire and Mello discovery was stuck at this problem-identification stage,

> "When people tried it and it worked, it was like ok let's work with it. Very few people thought it was worth studying, but everybody wanted to use it. So then you'd go to the worm meetings and everybody was using it." (respondent 12)

Problem-Solving Failures

Paradigmatic Pressures

For those who did see the phenomenon of gene silencing as a scientific endeavor worthy of pursuit, another barrier to breakthrough discovery from pursuing normal science is that scientists were constrained by current dogma when called upon to interpret unexpected results that often did not fit within the confines of current theory. Scientists take great pains in ensuring that the results they present are correct (at least within the state of knowledge and experimental

techniques available at the time), instead of risking the publication of artifacts that would eventually be refuted. To avoid being wrong when faced with weird results, they often chose to ignore anomalies and dismissed them as artifacts instead of proposing drastically different or even contradictory theories. This avoidance in challenging established dogma wasted valuable breakthrough discovery opportunities as scientists were confined by social and institutional norms of science that triggered the cognitive barrier of being pressured by paradigms.

The difficulty in explaining the observed silencing phenomenon and identifying the causal agent stemmed from a disparity in causal pathways between the RNAi mechanism and the central dogma of molecular biology. No one believed that dsRNA would work better than antisense RNA because if you had injected an antisense it would have immediately found its target whereas dsRNA would have to first unzip. Thus, scientists had to get passed being encultured (Simonton, 1989) by the doctrine that contradicted the ability for both sense and antisense RNA strands as well as dsRNA to perform equally well in silencing gene expression. This belief reinforced the established paradigm from the central dogma, which in turn constrained scientists from interpreting their results using a revolutionary framework even when a weird and interesting problem had been identified and pursued as a path of inquiry. Given the state of knowledge at the time, two informants illustrated how implausible dsRNA was to be the trigger to RNAi.

> "Nobody would ever inject the sense strand cause psychologically you could imagine how the antisense strand could work with the base pairing but the sense didn't make sense even though they showed they both worked equally well. No one ever did the sense strand cause they just thought that just can't be right. They just kind of ignored it and thought it's antisense." (respondent 1)

> "It's weird and not expected because basically we all knew that we make dsRNA and that's a dead end, it's an inhibition of the other RNA, you can't use that to make something." (respondent 15)

Boundary Barriers and Interplay with Paradigmatic Pressures

Boundary barriers in the problem-solving phase stemmed from scientists' belief of how fundamental the phenomenon of gene silencing traced back to a common ancestor between animals and plants. As an informant explained,

> "We know that plants evolved as a multicellular life forms independently from animals, so the last common ancestor of plants and animals was a single cell organism. And so when you're talking about how the cells are organized and develop, that happened independently. […] So when you're talking about very fundamental processes that were there in the last ancestor, last single cell ancestor, those operate across kingdom. So in general it just depends on whether you think it's an ancestral process or whether you think it's more derived." (respondent 3)

For those who saw the scientific merit of pursuing research on gene silencing, boundary barriers coupled with paradigmatic pressures reinforced each other in contributing to the failure of proposing a breakthrough, as illustrated in the bottom half of Figure 4. When faced with unexpected results in one single research setting and unable to gain more confidence from similar results in other settings due to boundary barriers, a scientist's ability to think in a revolutionary manner is compromised while being locked in the same mindset. Analogously, if one is constrained by current dogma and does not consider the possibility of a groundbreaking perspective, substantiation in other organisms and fields will not be sought out. In both cases, crisis will be missed and breakthroughs overlooked or delayed. An informant described this reticence that if one is alone in finding a contrarian result it is difficult to muster the courage to submit it for publication without having substantiated it somewhere else.

> "Cause if you think about it if you were sitting in a lab in the middle of nowhere injecting dsRNA into c. elegans, and seeing it having an effect, a really good effect, a really strong effect on gene expression and it doesn't work with single-stranded RNA, and no one has ever seen this before, you can't write this up. You must have put out a few fingers to see, whether anyone have heard of anything before." (respondent 7)

Unlike the animal community that was caught in the problem identification phase, most of the plant community was trapped in this problem-solving stage,

> "Why didn't the plant people get to where Fire & Mello did? My main insight is that we were so focused on transgenes to manipulate DNA expression. We never got to introducing RNA, it was regarded as unstable – that was never going to work. […] So there are these different mindsets that are so ingrained that you don't even appreciate that there is another way to look at this. And I think that's why we were really locked into that. It traced back to the discovery of DNA and the genetic material and the structure of DNA." (respondent 17)

Success in the Discovery of the First RNAi Causal Trigger

So why did Fire and Mello surmount the barriers to identify the opportunity for potential breakthrough? First, although Fire and Mello also first came in contact with the phenomenon from a tools development perspective when trying to inactivate genes using antisense oligonucleotide technology, they quickly realized that the phenomenon itself was interesting, important and worth studying. Instead of dismissing it as just a useful tool or a mere worm oddity they believed that it was a fundamental process conserved in other organisms. The following quotes exemplify both Andy Fire and Craig Mello's motivation to study the phenomenon from the point of view of their colleagues.

> "[Andy Fire] has always […] said look I think I can figure this out, and sometimes it's boring stuff, but he just latches on and keeps going." (respondent 12)

> "Craig [Mello] was very excited about it and he just wanted to figure it out. He thought it was fundamental, and he was right. He believed […] that if he figured it out, he would have done something good." (respondent 10)

Second, proposing dsRNA as the trigger agent in the seminal paper shows that Fire and Mello were able to see passed its inertness that they were taught to believe. Putting forth their

theory of RNA interference was an indication that they circumvented the constraints established by current dogma and were not encultured in the current thinking of their fields.

And finally throughout the discovery process, Fire and Mello were well aware of the work done by plant scientists and were able to connect the dots between these works, those from the c. elegans community and the results that they observed from their own experiments. From the Nobel paper citations which made reference to several related articles in plants, Fire and Mello were not only aware of the phenomenon in plants they also believed that it was similar to what they had discovered in worms.

Fire and Mello's success which hinged on surmounting all three barriers as necessary conditions to breakthrough discovery provides further evidence that these barriers cannot be viewed independently, but are rather interconnected and interact with one another.

DISCUSSION

The unifying theme that emanates from my findings is that at different stages of the discovery process scientists on the verge of discovery failed to identify or propose the breakthrough opportunity. This failure is based on a cognitive process triggered by institutional factors stemming from barriers at the problem identification and solving stages of the creative process. Besides contributing to the literature by showing and proposing a framework by which the seminal discovery was missed several times, scientists also hinted at a collection of practices they used to remedy the barriers and increase their likelihood of breakthrough. These practices can be operationalized as testable sources of breakthrough, but it is important to note that many of them are necessary but not sufficient.

For instance, to circumvent framing barriers scientists explored at the fringe while exploiting their focal research direction. This behavior can be proxied from the frequency distribution of scientist's MeSH keywords, where individuals who have a tendency to try high-risk explorative projects on the fringe are characterized by having a set of high frequency MeSH keywords representing the core of their research while at the same time having many one-off MeSH keywords mimicking the explorative nature of side projects. To avoid being confined paradigmatically, intra and inter laboratory collaborative ties used as a substantiation mechanism can be captured by the mix of organism specific MeSH keywords of papers written by a single author. If a scientist is associated with multiple organismic models from their published works' MeSH keywords, they have either worked with other labs that use different model organisms or run a lab that supports research in multiple organisms. To avoid boundary barriers besides turning to the literature, scientists also use conference attendance and teaching as ways to broaden their awareness in related fields. The role of conferences has been understudied in the literature but informants have suggested that the number of conferences and the breadth of conferences, whether interdisciplinary or cross-organism an individual attends is important to take into account as a source of breakthrough. Conferences are important not only as a perturbation to boundaries between communities of science to gain diversity of opinions and knowledge, but also act as a mechanism for result validation and provide a glimpse of informal scientific networks not captured through pure co-author collaborations. The number of cross-disciplinary courses a scientist teaches can also serve as a proxy for a source of breakthrough. Furthermore, another measure that proxies the scope of awareness of related research communities is the breadth of backward citations that scientists reference in their own

publications. This measure is also one of few bibliometric measures that can capture cognitive processes assuming that scientists cite papers that they are aware of.

This work informs the micro-foundations of the innovations literature by bringing individual level data to a question typically focused on the publication, patent or organization as the unit of analysis, or remained mainly theoretical. The literature in innovation has thus far mostly assumed constant input to innovation. My results suggest, instead, that individual inputs are quite heterogeneous and should be accounted for. Indeed, scientists are diverse and behave differently with regard to conference attendance, teaching, taste for exploitation versus exploration, collaborative preference and willingness to take closer steps or further leaps in research.

This work also solely focuses on cognitive barriers to breakthrough. Without doubt, intermixed with the barrier of framing the puzzling phenomenon as a tool rather than a subject of inquiry is the incentive pressure of consistently producing publications. Although the institution of science is based on the priority-reward system where participants are recognized for being first to discovery that pushes scientists to take on high risk and high rewards projects, they also face the realistic pressure of producing a steady stream of papers to satisfy funding conditions unless they benefit from sources that tolerate early failure, reward long-term success, and give its appointees great freedom to experiment (Azoulay, Graff Zivin, & Manso, 2011). The funding of research grants and the evaluation and promotion process within academia all play significant roles in determining the research path that scientists take. Compared to cognitive barriers these economic incentives, however, are more deliberate.

The other puzzle around RNAi centers on the fact that its initial use and perception as a tool did not facilitate discovery of its trigger mechanism but rather delayed it. RNAi is a perfect

illustration of the tension between concepts and tools because it effectively embodies both. Historians of science have extensively explored the two, and described how scientific revolutions arise from each. Kuhn (1962) perceived science from the point of view of a theoretical physicist, thereby emphasizing the great leaps of theoretical and conceptual insight that give rise to scientific revolutions for understanding nature while taking for granted experimental data. Whereas Galison's (1997) argument that new tools drive the process of scientific discovery stems from an experimental physics viewpoint where he described great leaps of practical ingenuity for observing nature enabled by the acquisition of new data. RNAi, however, is a hybrid that does not fit squarely in one camp or another. Instead, it is a tool based on an underlying biological concept. The case of RNAi suggests that the nature of the underlying knowledge should be a continuum rather than distinct categories between concepts and tools. RNAi debuted as a tool that arose from observations in plants, fungi and worms, but not understanding the causal mechanism to the phenomenon impeded its stability as a technique and consequently its initial widespread use and diffusion. It was not until the trigger agent was identified that the community of scientists started to study the intricacies of its mechanism in many organisms and RNAi became truly revolutionary. However, although in the beginning its perception as a tool delayed understanding of the concept, it promoted its diffusion once the trigger mechanism was understood. Familiarity brought about by its use eased acceptance of the underlying concepts.

Weaknesses and Limitations

As this work is a case study of one particular breakthrough, the main limitation is the generalizability of findings especially when studying idiosyncratic rare events like

breakthroughs. However, the goal of this work is to enhance the current understanding of how breakthroughs emerge with the counterfactual perpective of missed or delayed opportunities, and generate new theory on sources that enhance breakthrough potential that can be operationalized and broadly tested quantitatively for generalization. These include expanding research scope through exploration at the fringe to avoid being blinded by conventional science, exploiting social ties as a mechanism of substantiation to overcome being constrained by current dogma, and broadening exposure and awareness of work across multiple scientific communities to mitigate the inability to connect the dots.

Finally, any study that attempts to predict the source of an innovation will be sensitive to the definition of those at risk of innovating. My definition of the community of scientists attempting to solve the puzzling mechanism of gene silencing is mainly functional, but because the same phenomenon was named differently between plant, fungi and animal scientists my definition is by force of association also organism-based. This definition, however, is noisy given that the MeSH keywords I use are also assigned to other fields studying various biological phenomena, such as the interferon community. Other definitions of the community could have taken a purely model organism view rather than the phenomenon-based angle, whereby scientists in the plant, fungi and worm fields – the three communities that initially observed gene silencing – would make up the sample. However, this alternative would result in an even noisier set with the combination of three large organism-based communities regardless of the biological phenomenon each scientist focuses on.

CONCLUSION

The core findings of this paper complement our current understanding of how breakthroughs emerge by shedding light on a cognitive framework where the seminal discovery was missed several times because of failures to identify and propose the breakthrough opportunity because of three barriers. During problem identification, path dependence from established technologies and the quest toward normal science blinded scientists from recognizing a prospective breakthrough. They framed RNAi as a tool while ignoring it as a scientific concept worthy of study. Existing boundary barrier between communities of scientists prevented links from being recognized between several prior instances of odd observations and aggravated the difficulty in identifying the breakthrough opportunity by misrepresenting the magnitude of the problem. During problem-solving, scientists suffered from the socio-cognitive barrier of being constrained by current dogma. Due to fear of being wrong, they hesitated to propose solutions that significantly differed from established theory. Coupled with boundary barriers, similar anti-dogmatic results stayed isolated and diminished scientists' confidence to propose a new revolutionary paradigm.

This work has implications in the design of organizations and institutions that partake in scientific discovery. Understanding the barriers to scientific knowledge creation is vital not only for academic administrators but also crucial from both managerial and policy standpoints. It illustrates the fundamental differences inherent in the production of scientific and technological knowledge, and directly speaks to the organizational design of science-based firms (where the literature has mainly focused on technological innovation and remains thin) by providing structural characteristics and policies that foster the production of groundbreaking discoveries. These include facilitating interdisciplinary research teams, encouraging cross-organism and

cross-field conference attendance, and providing incentives that enable the flexibility to take on side projects at the fringe. From a policy vantage point, this work characterizes which scientists have the highest potential of breakthrough. This is a first step in eventually moving up levels of analysis to locating communities of scientists more likely to discover breakthroughs and, thus, enabling more targeted governmental subsidies and private investments into them (Lane, 2009).

A natural extension to the current work is to further quantitative operationalization and test sources that enhance breakthroughs uncovered herein. Keeping in mind tradeoffs in the practices scientists employ to circumvent barriers to breakthrough and dynamics between each theme, quantitatively testing these new sources of breakthroughs can shed light on equilibrium points as well as interaction effects. Another extension follows from the conventional wisdom of 'having smart people at the right place at the right time' when eliciting about breakthroughs. Therefore, in future work the question of when is the right time for breakthroughs to emerge can be studied. For instance, at what point in the maturity of a field are breakthroughs most likely to be made, what role do complementary discoveries – for instance microRNA and genome projects in the case of RNAi – play in spurring or stunting revolutionary discoveries, how much of an installed base is required within a community for breakthroughs to emerge are all intriguing questions to further explore.

The results from the predictive regressions that combine all existing sources of creativity and breakthrough are sobering, especially in attempting to predict the authorship of the actual Nobel Prize winning paper, as a collection of current theories of creativity, and scientists' prior publication productivity and quality, together can explain less than 1 percent of the model variance. Less stringent definitions of breakthrough were more successful, with almost 50 percent of the variance being explained. Such results can be seen pessimistically – we have

made essentially no bibliometric progress in predicting breakthroughs – or optimistically – we can predict half the variance in simple publishing productivity within a given field, given the history of the field.

From a policy standpoint, these results should give pause to the current efforts to use big data and computation to understand, justify, and optimize public investment in science. Policy makers and corporate lab managers should absolutely not apply bibliometric results blindly, given the large unexplained variance in the predictive regressions. Automated tools could certainly support a process run by domain experts. For example, as part of the peer review of grant applications, the predictive number could be calculated, but hopefully not over-interpreted, lest we kill (what still appears by all accounts to be the) golden goose of science. The current and typical peer-reviewed grant process may be inefficient and frustrating, but it is probably the least-worst method; it would be foolish to abandon it in favor of purely bibliometric criteria.

REFERENCES

Azoulay, P., Graff Zivin, J. S., & Manso, G. 2011. Incentives and creativity: evidence from the academic life sciences. ***The RAND Journal of Economics***, 42(3): 527-554.

Berson, J. A. 1992. Discoveries missed, discoveries made: creativity, influence, and fame in chemistry. ***Tetrahedron***, 48(1): 3-17.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. 1987. ***The Social construction of technological systems : new directions in the sociology and history of technology***. Cambridge, Mass.: MIT Press.

Burt, R. S. 2004. Structural Holes and Good Ideas. ***The American Journal of Sociology***, 110(2): 349-399.

Cockburn, I. M., & Henderson, R. M. 1998. Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery. ***The Journal of Industrial Economics***, 46(2): 157-182.

Cohen, W. M., & Levinthal, D. A. 1990. Absorptive Capacity: A New Perspective on Learning and Innovation. ***Administrative Science Quarterly***, 35(1): 128-152.

Collins, R. 1998. ***The sociology of philosophies : a global theory of intellectual change***. Cambridge, Mass.: Belknap Press of Harvard University Press.

Corbin, J. M., & Strauss, A. L. 2008. ***Basics of qualitative research : techniques and procedures for developing grounded theory*** (3rd ed.). Los Angeles, Calif.: Sage Publications, Inc.

Dasgupta, P., & David, P. A. 1994. Toward a new economics of science. ***Research Policy***, 23(5): 487-521.

Dosi, G. 1982. Technological paradigms and technological trajectories : A suggested interpretation of the determinants and directions of technical change. ***Research Policy***, 11(3): 147-162.

Dougherty, D. 1992. Interpretive Barriers to Successful Product Innovation in Large Firms. ***Organization Science***, 3(2): 179-202.

Dyson, F. J. 1972. Missed Opportunities. ***Bulletin (New Series) of the American Mathematical Society***, 78(5): 635-652.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. ***Nature***, 391(6669): 806.

Fire, A. Z. 2007. Gene silencing by double-stranded RNA. ***Cell Death & Differentiation***, 14(12): 1998-2012.

Fleming, L. 2001. Recombinant Uncertainty in Technological Search. ***Management Science***, 47(1): 117-132.

Fleming, L. 2002. Finding the organizational sources of technological breakthroughs: the story of Hewlett-Packard's thermal ink-jet. ***Industrial & Corporate Change***, 11(5): 1059-1084.

Fleming, L., & Sorenson, O. 2004. Science as a Map in Technological Search. ***Strategic Management Journal***, 25(8/9): 909-928.

Fortunato, S. 2010. Community detection in graphs. ***Physics Reports***, 486(3-5): 75-174.

Galison, P. 1997. ***Image and logic : a material culture of microphysics***. Chicago: University of Chicago Press.

Garud, R., & Rappa, M. A. 1994. A Socio-Cognitive Model of Technology Evolution: The Case of Cochlear Implants. ***Organization Science***, 5(3): 344-362.

Gieryn, T. F., & Hirsh, R. F. 1983. Marginality and Innovation in Science. ***Social Studies of Science***, 13(1): 87-106.

Gilbert, W. 1986. Origin of life: The RNA world. ***Nature***, 319(6055): 618-618.

Girotra, K., Terwiesch, C., & Ulrich, K. T. 2010. Idea Generation and the Quality of the Best Idea. ***Management Science***, 56(4): 591-605.

Golden, B. R. 1992. The Past Is the Past--Or Is It? The Use of Retrospective Accounts as Indicators of past Strategy. ***The Academy of Management Journal***, 35(4): 848-860.
Guo, S., & Kemphues, K. J. 1995. par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. ***Cell***, 81(4): 611-620.
Henderson, R. M., & Clark, K. B. 1990. Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. ***Administrative Science Quarterly***, 35(1): 9-30.
Jeppesen, L. B., & Lakhani, K. R. 2010. Marginality and Problem Solving Effectiveness in Broadcast Research. ***Organization Science***, Forthcoming.
Jones, B. F. 2009. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? ***Review of Economic Studies***, 76(1): 283-317.
Kaplan, S., & Vakili, K. 2012. Breakthrough innovations: Using topic modeling to distinguish the cognitive from the economic. ***Rotman School Working Paper.***
King, G., & Zeng, L. 1999a. Logistic Regression in Rare Events Data. ***Department of Government, Harvard University***.
Kogut, B., & Zander, U. 1992. Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. ***Organization Science***, 3(3): 383-397.
Krol, A. R. v. d., Leon, A. M., Beld, M., Mol, J. N. M., & Stuitje, A. R. 1990. Flavonoid Genes in Petunia: Addition of a Limited Number of Gene Copies May Lead to a Suppression of Gene Expression. ***The Plant Cell***, 2(4): 291-299.
Kuhn, T. S. 1962. ***The structure of scientific revolutions***. Chicago: University of Chicago Press.
Lane, J. 2009. Assessing the Impact of Science Funding. ***Science*** , 324: 1273-1275.
Leonard-Barton, D., & Swap, W. C. 1999. ***When sparks fly : igniting creativity in groups***. Boston, Mass.: Harvard Business School Press.
McEvily, B., & Zaheer, A. 1999. Bridging Ties: A Source of Firm Heterogeneity in Competitive Capabilities. ***Strategic Management Journal***, 20(12): 1133-1156.
McFadyen, M. A., & Cannella, A. A. J. 2004. Social Capital and Knowledge Creation: Diminishing Returns of the Number and Strength of Exchange Relationships ***Academy of Management Journal***, 47(5): 735-746.
McFadyen, M. A., Semadeni, M., & Cannella, J. A. A. 2009. Value of Strong Ties to Disconnected Others: Examining Knowledge Creation in Biomedicine. ***Organization Science***, 20(3): 552-564.
Merton, R. K. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science. ***American Sociological Review***, 22(6): 635-659.
Miles, M. B., & Huberman, A. M. 1984. ***Qualitative data analysis : a sourcebook of new methods***. Beverly Hills: Sage Publicaions.
Murray, F. 2002. Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. ***Research Policy***, 31(8-9): 1389-1403.
Napoli, C., Lemieux, C., & Jorgensen, R. 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. ***The Plant Cell***, 2(4): 279-289.
Obstfeld, D. 2005. Social Networks, the Tertius Iungens Orientation, and Involvement in Innovation. ***Administrative Science Quarterly***, 50(1): 100-130.
Ratcliff, F., Harrison, B. D., & Baulcombe, D. C. 1997. A Similarity Between Viral Defense and Gene Silencing in Plants. ***Science***, 276(5318): 1558-1560.
Rivkin, J. W., & Siggelkow, N. 2002. Organizational sticking points on NK Landscapes. ***Complexity***, 7(5): 31-43.

Romano, N., & Macino, G. 1992. Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. ***Molecular Microbiology***, 6(22): 3343-3353.

Schumpeter, J. A. 1934. ***The theory of economic development; an inquiry into profits, capital, credit, interest, and the business cycle***. Cambridge, Mass.: Harvard University Press.

Schumpeter, J. A. 1942. ***Capitalism, socialism, and democracy*** (1st Harper Perennial Modern Thought ed.). New York: HarperPerennial.

Simonton, D. K. 1989. Age and creative productivity: Nonlinear estimation of an information-processing model. ***International Journal of Aging and Human Development***, 29: 23-37.

Simonton, D. K. 1999. ***Origins of genius : Darwinian perspectives on creativity***. New York: Oxford University Press.

Singh, J., & Fleming, L. 2010. Lone Inventors as Sources of Breakthroughs: Myth or Reality? ***Management Science***, 56(1): 41-56.

Stern, S. 2004. Do Scientists Pay to Be Scientists? ***Management Science***, 50(6): 835-853.

Swanson, D. R., Smalheiser, N. R., & Torvik, V. I. 2006. Ranking indirect connections in literature-based discovery: The role of medical subject headings. ***Journal of the American Society for Information Science & Technology***, 57(11): 1427-1439.

Teece, D. J., Pisano, G., & Shuen, A. 1997. Dynamic Capabilities and Strategic Management. ***Strategic Management Journal***, 18(7): 509-533.

Thomke, S. H. 2003. ***Experimentation matters : unlocking the potential of new technologies for innovation***. Boston Mass.: Harvard Business School Press.

Torvik, V. I., & Smalheiser, N. R. 2009. Author Name Disambiguation in MEDLINE. ***ACM transactions on knowledge discovery from data***, 3(3): 1-29.

Uzzi, B. 1997. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. ***Administrative Science Quarterly***, 42(1): 35-67.

Waterhouse, P. M., Graham, M. W., & Wang, M.-B. 1998. Virus Resistance and Gene Silencing in Plants can be Induced by Simultaneous Expression of Sense and Antisense RNA. ***Proceedings of the National Academy of Sciences of the United States of America***, 95(23): 13959-13964.

Wuchty, S., Jones, B. F., & Uzzi, B. 2007. The Increasing Dominance of Teams in Production of Knowledge. ***Science***, 316(5827): 1036-1039.

Figure 1 – Effect size of each regression model by explanatory variable. The effect size is computed by increasing the variable under study by one standard deviation from the mean while holding all other explanatory variables at the mean.

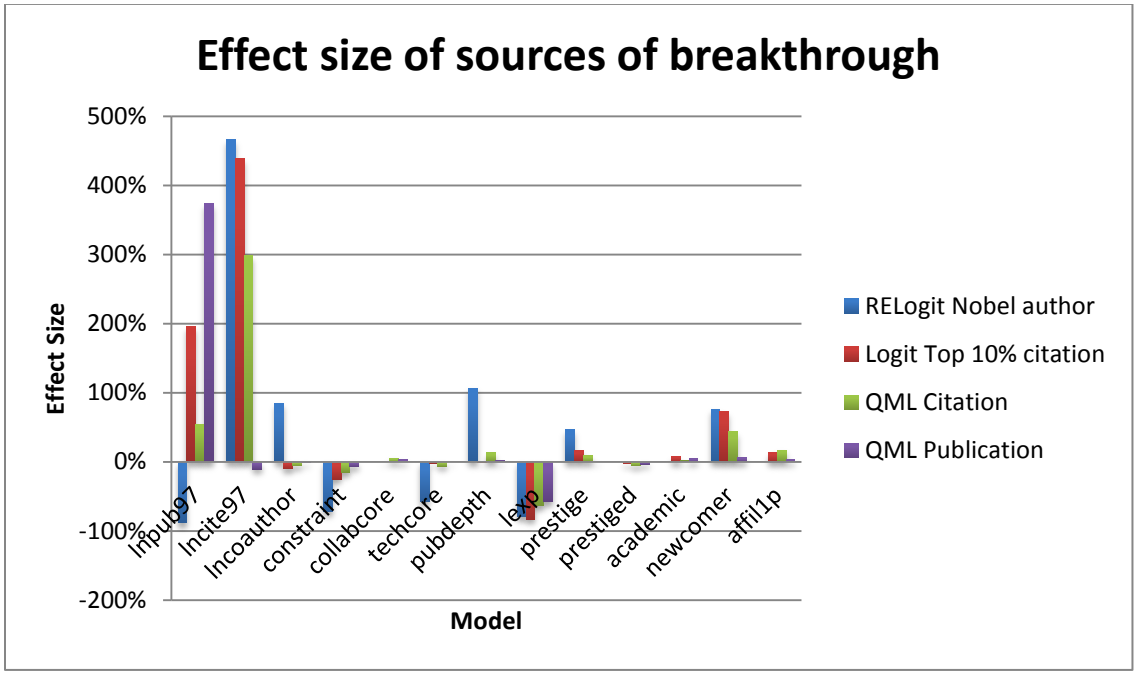

Figure 2 – Explained variance for each model and source of breakthrough explored

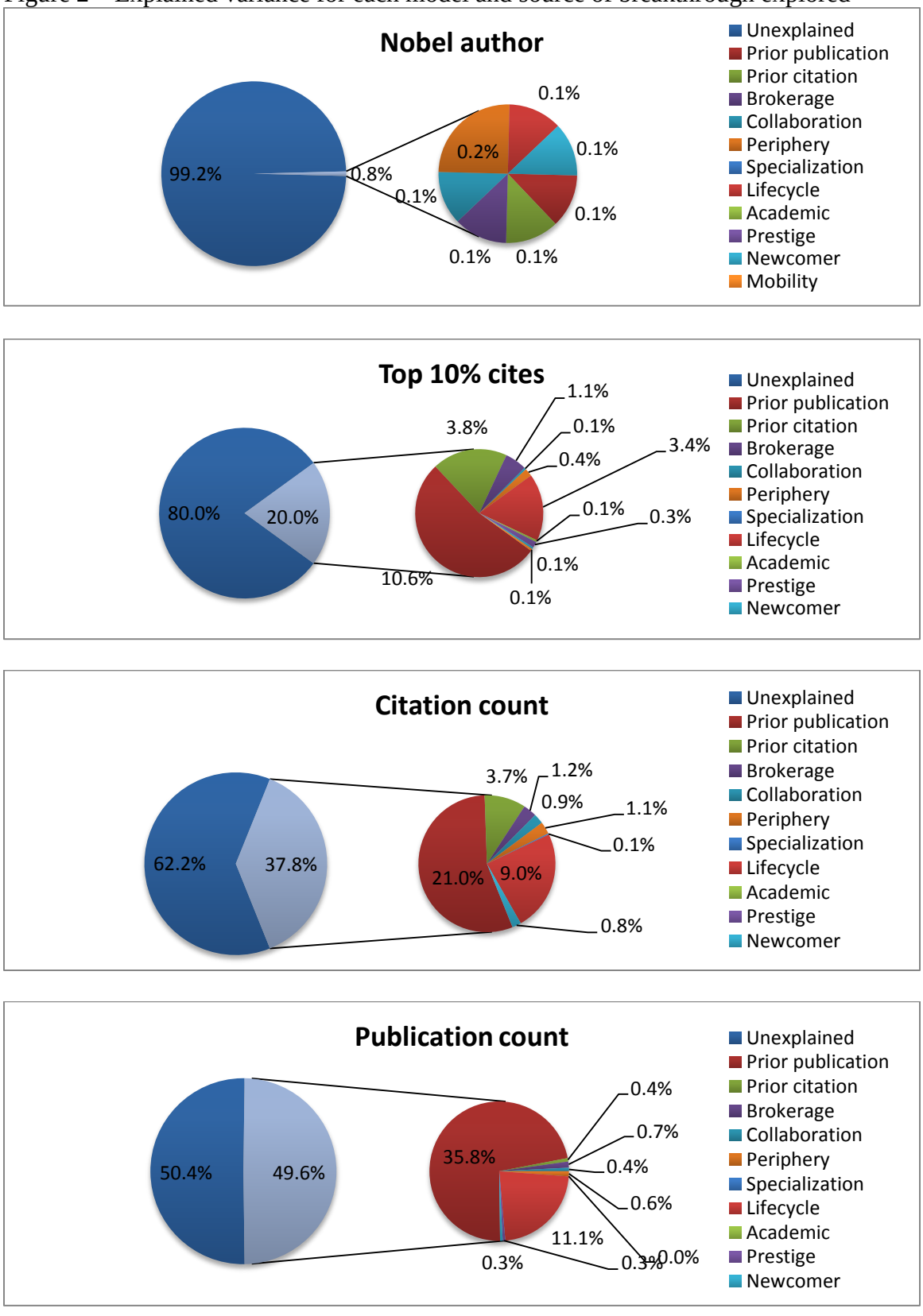

Figure 3 – Network plot of RNAi community with nodes representing each scientist, link as co-authorship relationships. Blue represents actual impact and pink represents predicted impact.

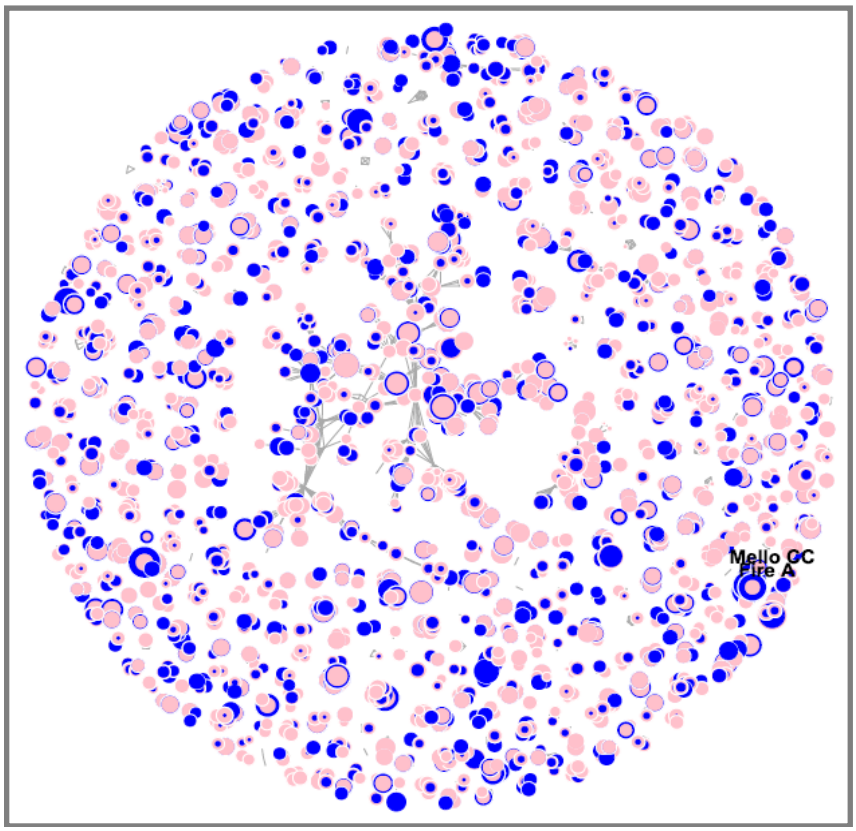

Figure 4 – Framework of Missed Breakthrough

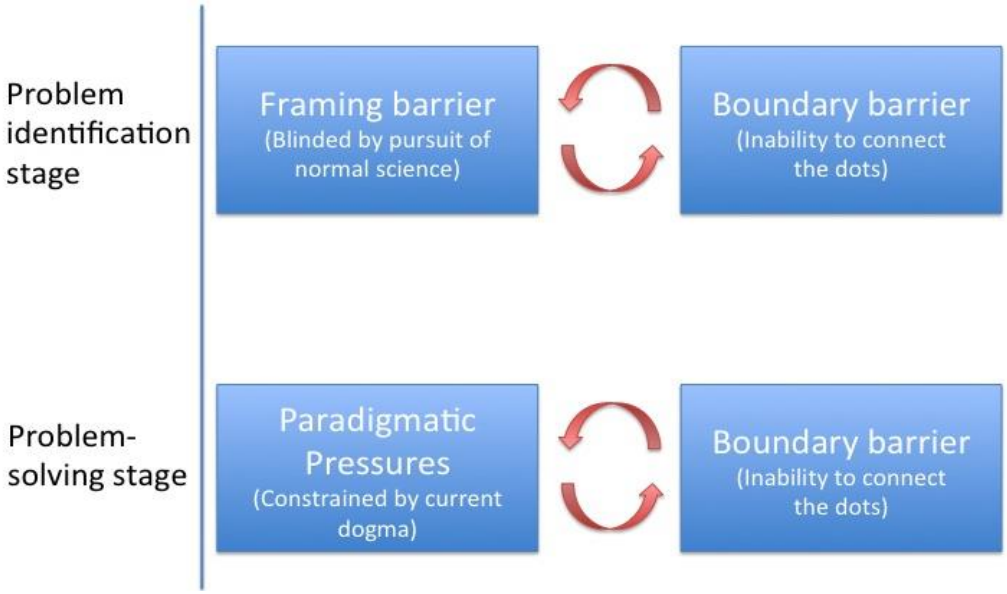

Table 1 – Informant quotations respectively illustrating framing barriers, paradigmatic pressures and boundary barriers that contributed to the delay of the RNAi breakthrough discovery.

	Model Organism	Framing Barriers	Paradigmatic Pressures	Boundary Barriers
1	c. elegans	"It's really puzzling, people had just filed this away and just thought this doesn't make sense but didn't think about it. They just kind of ignored it and just thought it's antisense, people used to call it antisense even though it wasn't."	"It was funny because the sense and the antisense strand both worked. Nobody would ever inject the sense strand cause psychologically you could imagine how the antisense strand could work with the base-pairing but the sense didn't make sense even though they showed they both worked equally well." "People talk themselves out of doing experiments all the time. They'll say I won't try that because it will probably look like this. Maybe. Maybe not! If you don't do it you never will know."	"No one at that time, no one I had talked to was even thinking that it was related to the plant things. No one before 98 I had ever heard anyone mention anything to do with plants."
2	c. elegans	"We were obviously intrigued by it, but we could use to probe some biology that we were interested in it. And you want to do in science, it's almost like you see something and you want to harvest it. So we could harvest RNAi in a way by using it as a novel method, it allows you to leverage some biology. You didn't have to get mutations and you could get some information and learn something about it. The community started to adopt it as a method, because they knew it was specific."	"It's hard to do these breakthroughs where you really have to step beyond your comfort zone."	"We were trying to penetrate what we thought what we thought was a novel phenomenon. We didn't believe that it really represented anything general." "If you have, if your suspicion let's say the weight of your suspicions is that it's probably kind of worm specific, what's the point of devoting a lot of resources to it. Because we are trying to figure out things that are general and broad right. So the fact that the worm could do this and that other things couldn't do it. I mean flies don't do it, and it's inconceivable that mammals would do it. So you're thinking it's a worm thing."
3	drosophila	"Scientifically you know that this is working and these people were just using this as a tool. Then you have to decide ok. On the one hand, this is just a tool and the reason you're using this tool is because you want to study the biology of these genes and you're really focused on that biology and so you're convinced that using this antisense method is teaching about the function of those genes and you go on and you focus on the function of those genes. And, you don't get distracted by this oddity that the sense is also working."		"I didn't know that the plant phenomenon would be related to the worm phenomenon. Obviously the people working in plants, were in fact trying to explain the same phenomenon but we didn't know any of those details."

4	plants	"I mean a lot of people have been doing the sorts of experiments, putting genes in viruses into plants. A lot of people had been seeing exactly the same thing, they had seen that the transgene that conferred resistance was not expressed that it was silenced. And they just ignored it. So people sometimes ignore data when it stares them in the face."	"[In plants] it was largely phenomenology, there wasn't a lot do to with the mechanism." "However we had done some experiments that implicated dsRNA. We had done 2 experiments one set of which never got published. We did submit them for publication and then they came back and the editor said they were of insufficient general interest. And so the reviewers were not convinced that, they thought that it's an interesting illustration of how a field can get preconceptions."	"We were aware of the worm story to some very small extent, partly because what we knew was Ken Kemphues' original micro injection experiments. So we knew about those and those looked like some sort of co-suppression phenomenon and also I knew about Ruvkun and Ambros' work. We missed the link between that work and our work, so the small temporal RNA, the Ruvkun and Ambros work so that looked as if it were a translational regulation thing." "I don't think there were really conferences that brought the animal and plant fields together until probably as late as 2001."
5	c. elegans	"People who were using antisense were using it to inhibit genes so they could show in vivo or address inactivity in a gene in vivo."		"I think the effects of antisense were not that satisfactory cause they were not that potent and hard to control for. So again it gets back to this question, yeah if you have a control that doesn't make a lot of sense, you are not going to report it. Cause there are probably a lot of observations that were not, experiments that were not actually included in papers. You know one of the problems with science is that negative results often do not go reported. And they are left un-described."
6	drosophila	"They just looked at this like a bizarre tool, they couldn't explain it but it was fabulous for what they wanted to do. They could silence genes, so it was kind of like this. They were focused on the thing at hand and kind of ignoring this elephant in the room, which was far more important and interesting."	"Craig got up to share with us that workshop in 97 RNAi. We thought this was really bizarre. I remember being there; everyone in the room was bedazzled, because I was so bizarre. I ran counterintuitive to everything we've been taught." "We knew [the experiment] had worked. It's like holy shit, although you're really scared that you're over interpreting it or something."	"The quelling people were kind of off on their own; they didn't interact very much with us." "There is this really bizarre phenomenon but it never occurred to me at that point that it would be applied to other organisms. So it never occurred to me at that time that I should try RNAi in drosophila."
7	c. elegans	"Everybody was wanting a way to knock a gene down. So it was a receptive community, when people saw that they thought that's very interesting. And I don't think it was long before everyone was trying it out. Everyone could see the application of it, everyone was already primed to apply it."	"And you know, a lot of people had described similar things or talked about it certain things. But nobody took it terribly seriously. Why should dsRNA work better than antisense. Because if you had sticked in an antisense providing it doesn't get degraded, it should find it's target and take it out. When you put in pre-annealed dsRNA we thought it had to unzip, it's actually a weakness, because nobody realized that there is a machinery that does that."	"[The plant section] might as well be lectured in Latin, they don't really integrate with other people. They didn't think animals did the same thing." "If you were sitting in a lab in the middle of nowhere injecting dsRNA into C. Elegans, and seeing it having a really strong effect on gene expression, it doesn't work with single-stranded RNA and no one has ever seen this before, I can't write this up. You must have put out a few fingers to see, whether anyone have heard of anything before."

8	c. elegans	"So I was also able to also use the technique to inhibit that gene activity and see."		"We didn't even know that, it would become such a general phenomenon. No one knows because we thought that it is something peculiar with worm. Actually I was going to discuss some of these [plant results] in my original paper but my advisor felt it was a little too premature to make that kind of link."
9	mammals	"So we used gene inhibition technologies in order to understand the pathways. But at the same time we also worked on developing this type of technologies, so I think I was more or less among the first, the ten first, to use antisense oligonucleotides which were very popular, but much less efficient than RNAi."	"Who is going to think let's put a double-stranded short RNA, if you don't know the system, who is going to say let's put a short double-stranded RNA in a cell by chance and it could be something. It's impossible. It was so anti-dogmatic, because there was DNA, RNA and protein. I guess it took time also and very bright and inventive people to really go against the dogma. And say ok, maybe something is wrong. It is always very difficult."	"At this time, I had never followed what was going on in plants. Preconception that whatever perhaps happens in plants is different in animals even in mammals, that there wasn't much attention paid to them, even within the community. The same thing happened with c. elegans in a way. Because at the beginning everybody thought, ok this thing is interesting but most of them thought just in c. elegans."
10	c. elegans	"And so we never asked the question in a serious way other than talking out of this work. So that we then continued to use the technique because it was clear that one of the key control experiment was to show that it wasn't any old RNA that did this effect so it was a very gene specific effect and so once we knew it was a gene specific effect we didn't really care how it worked. All we cared about was that we could use it. Everybody was very excited about it because of the potential for its use to target specific genes without going through the trouble of making mutations. People were intrigued but that's different from going after it."	"Because I didn't have that information, that knowledge. I wouldn't have made that connection it never occurred to me that there was both strands in our reaction."	"As far we knew at the time, it was a very specific phenomenon for c. elegans. It was pretty much just thought of as a c. elegans phenomenon at the time."
11	mammals	"Fire and Mello were trying to do an antisense. So people were using antisense oligonucleotides for a long time to try to do what RNAi does."	"dsRNA molecules were not typically viewed as being naturally occurring molecules. They were typically viewed as being part of virus or whatever."	"The paper by Jorgenson on petunias in plant cell, the name of the journal was Plant Sciences, Plant Cell or something. That really nobody followed, nobody. It's interesting the plant world tends to have its own group of people. And they don't tend to intermix too much with the non-plant people."

12	c. elegans	"Cause it's a tool that everybody wants to use like recombinant DNA, many people who wanted to use it don't care how it works it just becomes a tool that they use. When people tried it and it worked it was like ok, let's work with it. very few people thought it was worth studying. But everybody wanted to use it. So then you'd go to the worm meetings and everybody was using it."	"First reaction was: it can't be right, it's too weird."	"Rich Jorgenson and Carolyn Napoli, they were telling me stories about silencing they put in. They had these flower color things, trying to get purple it would turn white, it was all screwed up. But I missed it entirely. I did not see the connection."
13	drosophila	"But they were so focused upon the objective they were studying – was this gene required for this mutant phenotype. It confirmed the issue they designed the experiment for that was their objective. So they went and published the results, saying this is the function of this gene. But what they didn't do it that they didn't say that this control that didn't work is likely to be more important than this paper. And we should put aside the results of this paper and pursue that control."	"Yes, and there was DNA, RNA and protein. And then we started to get information from RNA to DNA. It started to be a big change that RNA also could be considered as an information and not just as an intermediary."	
14	plants	"You could silence genes, which was a tool that could be valuable."	"We are not really discovering a new thing, in fact what we are discovering are things that already existed but that we are simply ignoring or that we underestimated." "Also you discover it was something really different than the dogma that dsRNA played a role in animals."	"Here we are too isolated; we don't have enough interaction with people because we are only working on plants. We are experts on plants but we are only working on plants." "There was not that many meetings that mixed organisms." "So we extended this from plant to fungus, but still in 1996 there was nothing published on animals, at all."
15	e. coli	"Even though nobody knew really how it worked, the success rate was enormous. You could use it. And there were all these patents in the early. Because you could try sense or antisense, or you use a sense gene or antisense gene generally speaking it worked."	"Almost none of us thought it would be the discovery that dsRNA is the trigger, that is something we did not expect. It's weird, not expected because basically all we knew that we make dsRNA and that's a dead end, it's an inhibition of the other RNA, you can't use that to make something." "They didn't really know what to do with it cause dsRNA doesn't do anything."	"The people who should have picked up on this and Victor [Ambros]' discovery in the beginning they didn't because it was a worm thing, worms are doing strange things."

16	c. elegans	"Somebody took me aside and said whatever you do don't just work on RNAi, because it's not biology it's just a technique."	"The hypothesis that were successful for you in explaining phenomenon A, kind of get recycled as the first choice in explaining phenomenon B. Because it's what you're most comfortable with and you know how to test it and if it's wrong you now have this set of experiments that helps point you in the right direction." "So I was just fascinated with the idea that dsRNA could do anything, I had been brought up to believe it was inert."	"Because I have always thought that the real barrier to productivity in science was people not communicating immediately when they see the common thread. Once you wait until your discovery is polished and presented then you've already filtered out some of the things you don't understand that the right person can explain to you because they have the other piece of the puzzle."
17	plants	"You have a particular objective you want to understand X you want to solve that you're using hypothesis testing, I think that turns out to be kind of a trap."	"So there are these different mindsets that are so ingrained that you don't even appreciate that there is another way to look at this. And I think that's why we were really locked into that. It traced back to the discovery of DNA and the genetic material and the structure of DNA. […] We were manipulating DNA not RNA. That was the one missing piece, had we gone into introducing RNA directly we could have done things like Fire & Mello did and we could have done it years before them."	"We were publishing in different journals then a lot of the animal folks that yeast folks wouldn't see if they were at the wrong kind of institution, and that created an artificial barrier that doesn't exist now but was an important one then."
18	plants	"So I would say that in that case we were trying to do something different, this gene replacement, but in the process of doing those experiments we stumbled upon this gene silencing and at that point it was so interesting, it seemed so new and not explainable by anything that we had known before that we had started focusing on that phenomenon." "Everybody wanted to use this technology first as a technology for research for knocking down a gene."	"We were subconsciously ignoring a lot of science. We were also testing with what kind of thing do you need to trigger this gene silencing and we had already setup this experiment to test this that there would be some kind of RNA signal involved, and results at the time also suggested that it was likely to be dsRNA."	"There wasn't a lot of dialogue then between the plant and animal community. […] And at the time I don't think we were thinking too much about necessarily the animal work. But during the initial years when we were working on it I think we weren't talking with animal people very much, it was more just a small group of plant scientists who were first trying to figure out what was going on." "Plant scientists find that a lot of animal scientists don't take you very seriously. But there are so many fundamental biological findings made in plant systems beginning with Mendel and his peas and the genetics. But we sort of felt like we were on the side, the animal people would always listen to animal people."

APPENDICES

Appendix A: Detailed Methods and Results for the Predictive Regressions

Models

Given the scientists' bibliometric attributes before 1998, I predict who would have a fruitful year in 1998 thereby identifying individual sources of breakthrough. The explanatory variables consisted of measures calculated using each author's prior bibliometric data up to 1997 inclusively encompassing all papers available in the PubMed database, while publication data in 1998 are used to calculate outcome variables.

I first predict the breakthrough itself using rare events logistic (relogit) models on all authors of the Nobel paper, an extremely infrequent event with only six successes out of the entire sample of nearly four thousand scientists. The relogit regression is the same as a standard logistic regression but corrected for bias when observed events are rare. I used logistic models with cluster robust standard errors for scientists who are in the top ten percent of the citation distribution. Finally, I employed count models when the forward citation counts of 1998 papers and the number of publications in 1998 were outcome variables. The count models are quasi-maximum likelihood Poisson (QML Poisson) with robust standard errors since publications and citations are non-negative counts and over-dispersed which prevents the use of standard Poisson models where it is assumed that the mean and variance of the variable distribution are equal. I also ran OLS regression models to evaluate the predictive power of measures of sources of breakthrough.

I restricted the dataset used in the empirical analysis to the PubMed Author-ity database (Torvik & Smalheiser, 2009) since the breakthrough is in the life sciences, and organized each data point as unique author records. I extracted unique authors from the set of papers obtained in my identification of the community of RNAi scientists. This sample of scientists defining the pre-breakthrough RNAi community yielded 1,551 papers and 3,959 unique authors. Due to missing affiliation data for 49 such individuals, the sample shrunk to 3,910 authors. Out of the 3,959 unique authors present in the sample, 144 authors do not have any prior publications either within the RNAi community or any other tangential field within the life sciences. Including these newcomers in the sample illustrates the true dynamic nature and evolution of such scientific communities. However, it also complicates data collection since there is no historical information on them prior to 1998 and leaves several explanatory variables undefined. Undefined variables for newcomers were set to the mean.

Dependent Variables

Nobel paper dummy – nobeldum is an indicator that equals one for the six authors on the Nobel winning RNAi paper in 1998.

Top ten percent citations dummy – Measures of publication impact based on citations rest on a social definition of creative success, where scientists are only thought to be creative if they receive recognition from their community or society as a whole, and their work is used as a foundation for further advancements (Simonton, 1999). I therefore relaxed the definition of breakthrough from the Nobel paper to scientists with citations in the top ten percent of the citation distribution with the indicator top10cite. Dependent variables for the top five percent returned similar results.

Number of forward citations for 1998 publications – I further relaxed the operationalization of breakthrough using forward citation counts garnered until 2010 of 1998 publications (ncite98), which rests on the same premise of social construction of success. For the OLS models I take the natural

logarithm plus 1 (lncite98) to match count explanatory variables that underwent the same transformation due to the Poisson models.

Number of 1998 publications – The final dependent variable is a measure of productivity (npub98) depicted by the number of 1998 publications. Similarly, I also take the natural logarithm plus 1 (lnpub98) for the OLS regressions.

Explanatory Variables

Publication history and eminence – Publication history is the count of one's total number of publications since first publishing until the year prior to the 1998 breakthrough (npub97) while publication eminence is the number of aggregated forward citations to these publications (ncite97). When npub97 is zero the scientist does not have any prior publications and is a newcomer. When ncite97 is zero the scientist could either be a newcomer with no prior publications or could have produced prior publications that have not been subsequently cited. Since the quasi-maximum likelihood Poisson count model is employed in the regressions and both variables are counts, their natural logarithm is taken and denoted respectively as lnpub97 and lncite97.

Collaborative vs. Individual researchers – The number of co-authors (ncoauthor) each scientist has worked with was captured for all publications prior to 1998 by calculating the degree network measure of each author node – number of directly linked neighboring nodes to a focal node. The network is portrayed by collaborative co-authorship ties for all publications prior to 1998 for researchers within the RNAi community as defined in the prior section.

Brokerage vs. Cohesion – Using the same network depiction, cohesion was measured by calculating Burt's constraint (Burt, 2004). To calculate the constraint, $C_i = \sum_j c_{ij}$ where $c_{ij} = \left(p_{ij} + \sum_k p_{ik}p_{kj}\right)^2$ and p_{ij} is the fraction of i's relation invested in contact j. p_{ij} translates to the degree of i if there is no prior weight to the social networks or, in other words, all connections are considered to be equal strength. Since cohesion is undefined for newcomers, newcomers' cohesion was set at the average cohesion value without taking into account newcomers.

Periphery vs. Core – Two measures of core were created to mirror the topical and collaborative community. The first measure depicts collaborative core of the scientific community, where core is structurally operationalized by the indicator variable collabcore. Scientists situated in the largest connected component of the RNAi network with the most number of interlinked collaborations are considered to be in the core and hence assigned a value of one, while all other scientists including newcomers are considered to be in the periphery of the community and take on the value of zero.

The second measure depicted core versus periphery from a technical standpoint (techcore). Following the topical construction using MeSH keywords of the scientific community working on suppressing gene expression, technical core is calculated by tabulating the frequency of MeSH keywords used in the community's definition[1] and all previous variants[2] in a scientist's publication history and normalizing by the total frequency of all MeSH keywords, i.e. $\frac{\text{Mesh freq of RNA,dsRNA+RNA,Antisense+Gene Expression Regulation+Gene Silencing+RNA,catalytic}}{\sum_i \text{freq of MeSH}_i}$. The more a scientist's work is focused in the key antecedent fields to RNA interference, the more they are embedded in the technical core of the community. The dataset provides the top 20 most frequent MeSH keywords

[1] "RNA, Double-Stranded", "RNA, Antisense", "Gene Expression Regulation" and "Gene Silencing", "RNA, catalytic"

[2] Prior MeSH keywords for "Gene Expression Regulation" include "Gene Expression", "Genes" and "Phenotype". When tabulating frequency for "Gene Expression Regulation" I also incorporated counts of its prior keywords.

per author and so for many scientists in the sample the majority of their work is not in precursor fields to RNAi. Hence those who's top 20 most frequent MeSH keywords do not match none the above five MeSH keywords take on the value of zero for techcore. Moreover, as 55 percent of the values of this variable are zero, representing a non-core position, I dichotomize this variable. Any non-zero value of this variable takes on the indicator value.

Specialist vs. Generalist – It can be difficult to differentiate the notions of periphery and core with those of specialist and generalist. However, these concepts can be quite different. A researcher can be specialist in one field in which they possess deep expertise while simultaneously be at the periphery of another. Similarly nothing prevents a generalist to be situated at the core of a given community.

The degree of expertise of each individual scientist was depicted using a publication breadth measure (pubdepth) implemented based on the breadth of MeSH keywords in a scientist's publications. This metric is a measure of the prominence of high-frequency peaks in the unique list of MeSH keyword distribution by author. The top most frequent number of MeSH terms for each scientist, k[3], was first identified. Again with the top 20 MeSH keywords for each author, k=5, and publication depth was calculated as the ratio of the frequency sum of the top 2 to 6 most frequent MeSH keywords, i.e. the high frequency peaks, to the sum of the frequency of all MeSH keywords from range 2 to 20, $\text{pubdepth} = \frac{\text{sum of MeSH freq in range 2 to k+1}}{(\text{sum of MeSH freq in range 2 to k+1})+\text{sum of remaining MeSH freq}}$ (Swanson, Smalheiser, & Torvik, 2006). A specialist with a narrow range of high frequency MeSH keywords has a high value in the numerator, and consequently has higher depth values; whereas a generalist tends to be characterized by a more uniform set of MeSH keyword frequency distribution and less defined high-frequency peaks which translates into lower numerator and depth values. Since specialist is undefined for newcomers, newcomers' pubdepth were set at the average value without taking into account newcomers.

Lifecycle – Scientists' experience was proxied by the number of years since first publication. Newcomers have zero years of experience while seasoned scientists may have several decades under their belts. Due to the model specification and the count nature of the variable, I take the natural logarithm and denote as lexp. Non-parametric modeling of this variable supports use of the more parsimonious first degree logarithmic of the variable (its effect is increasingly and monotonically negative).

Organizational Affiliation – The proportion of a scientist's academic affiliations was stored in variable academic. Academic equals one for a pure academic scientist and zero for a scientist working strictly in industry. Therefore if a scientist has a total of 3 affiliations, 2 in academia and 1 in industry, their value for academic would be set to 2/3.

Prestige – Prestige of a scientist's affiliated institution is a weighted average score with weights assigned according to the top 50 overall research universities as ranked by U.S. News in 1998. The best university is assigned a weighted score of 50 decreasing to a score of 1 for the 50th ranked university, while institutions beyond 50 receive a score of zero. $\text{Prestige} = \frac{\sum_{i=0}^{n} \#\,\text{publications}_i \cdot \text{university score}_i}{\text{total \# publications}}$ where n is the total number of unique affiliations. Furthermore, I add an indicator of being affiliated with a top 50 ranked institution at least once (prestiged) so as to correct the skewed distribution with the above weighted measure of prestige. For newcomer both measures of prestige are set to zero.

Mobility – An indicator for mobility between organizations or institutions was turned on if the total number of one's affiliation is greater than one (affil1p). I also included a dummy variable for newcomer (newcomer) equal to one if a scientist appears in the sample only starting in 1998.

[3] *k =int(1.7ln(u)+0.5)* where *u*=number of unique MeSH for individual i = 20, so k = 5.

Results

Table A.1 shows the summary statistics and correlation matrix for all dependent and exploratory variables used throughout the regression analyses. Out of the 3,959 unique authors 144 are newcomers. Only one author from these 144 newcomers was affiliated with a non-academic institution. Furthermore, only one newcomer published as a solo author in 1998 while all others collaborated either as first author (n = 57), last author (n = 10) or appeared as a middle author (n = 93). These collaboration structures reflect the apprenticeship model for graduate studies in the life sciences. Because most of the publications in which new scientists partake are co-authored with the principal investigator and other lab members, the number of new authors getting cited at least once is relatively high (n = 127).

Table A.2 reports results for the models that predict authors of the Nobel winning paper, scientists in the top ten percent of the citation distribution, citation count for 1998 papers and publication count in 1998. The rare events logit (King & Zeng, 1999a) provides an estimate that a particular scientist is most likely to discover the Nobel winning breakthrough. My results show that specialized brokers with prior eminence but less publication history prevailed. The model drops academic since it is a perfect predictor of authoring in the Nobel paper. It also drops collabcore, prestige dummy and affil1p because these three variables are perfectly multicollinear to newcomer for the six authors of the Nobel paper.

The logit model assesses the probability of a scientist publishing impactful papers with total citation counts in the top tenth of the distribution. As expected a top ten percent citation scientist is positively and significantly associated with their prior publications' impact as the number of forward citations for pre-1998 papers. I also find evidence that increased prior productivity and brokerage significantly increase the likelihood of attaining this top tier in citation. No significant correlation is present between the number of co-authors, structural or social core, specialization, or multiple affiliations. A linear relationship is observed where younger scientists correlate significantly with high citation likelihood. While the proportion of academic affiliations shows no significant relationship, the prestige of the affiliation is positively but weakly significant. Newcomers positively affect the likelihood of being on top of the citation distribution. Calculating the effect size for significant variables, a one standard deviation increase from the mean increases 95.1 percent[4] for the natural log of prior publication count, 338 percent for the natural log of prior citation count, and 72.7 percent for newcomer; and decreases of 24.9 percent for constraint and 83.1 percent for natural log of experience to the probability of being in the top ten percent of citations. In sum, younger scientists with more prior history and eminence, and situated in brokerage positions of their field have a higher probability to be in top tenth of the citation distribution. Being a newcomer also contributes positively.

The third model shows results for the first QML Poisson regression with forward citation count as the outcome variable. The results of this model are similar to those found from the logistic model, as both dependent variables depict a similar concept of impact. Interpreting the effect sizes, a one standard deviation increase in the natural log of prior publication citations increases citation count by 198 percent[5], while a one standard deviation increase in the natural log of prior publication increases citation count by 54.3 percent. Similarly, the coefficients on constraint and log experience indicate that a one standard deviation increase in each of the two variables decreases citations by 15.4 percent and 63.3 percent respectively. A one standard deviation increase in newcomer increases citation count by 43.6 percent. In summary, I find that younger scientists with more prior eminence and history, situated in brokerage positions are more likely to discover breakthroughs.

---

[4] Effect size = $\frac{1+e^{-(\propto+\beta_i\cdot\mu_i)}}{1+e^{-(\propto+\beta_i\cdot(\mu_i+\sigma_i))}}-1=\frac{1+e^{-(\propto+\beta_{inpub97}\cdot\mu_{lnpub97})}}{1+e^{-(\propto+\beta_{lnpub97}\cdot(\mu_{lnpub97}+\sigma_{lnpub97}))}}-1$

[5] Effect size = $\frac{e^{\beta_i\cdot(\mu_i+\sigma_i)}}{e^{\beta_i\cdot(\mu_i)}}-1=\frac{e^{\beta_{lncite97}\cdot(\mu_{lncite97}+\sigma_{lncite97})}}{e^{\beta_{lncite97}\cdot(\mu_{lncite97})}}-1$

The fourth model presents the regression with the measure of productivity proxied by the number of papers published in 1998 as the dependent variable. Prior publication quantity is positively and significantly associated with productivity, whereas contrary to models 2 and 3 that depict impact prior publication quality contributes negatively to productivity. Scientists who prioritize quality over quantity are less productive in order to ensure the quality of their work. Brokers with their nexus positions are more productive, whereas the number of co-authors is still insignificant. Again no significant evidence of periphery or core nor specialization or generalization is observed. I also find that younger scientists write higher impact papers and more papers, possibly because they are incentivized by the tenure system at most academic institutions. Scientists in academic institutions are more productive. Finally, newcomer also positively affects publication count.

I use OLS to shed light on the predictive power of the theoretical models. Figure 2 shows the percent contribution to the variance of each explored theory for all four models. In these models I only interpret R^2 and delta R^2 measures even though coefficients are directionally consistent with the more appropriate non-linear models. Unsurprisingly, predicting who will discover the breakthrough per se is extremely hard due to the rare nature of such events. Any model trying to predict 6 successful events out of 4,000 would be challenged, and this is illustrated by the total R^2 of 0.8 percent when predicting authors of the Nobel winning paper (regression table available from author upon request). The total explained variance jumps to almost 20 percent, there is still only have one out of five chances of predicting the correct top ten percent citation scientist. But without prior publication and eminence, prediction using the theoretical themes only provides 5.6 percent explained variance. Further relaxing predictive requirements to the number of citations yield increased total explained variance to 37.8 percent, where together measures of prior eminence and productivity account for 24.7 percent of the variance and the remaining theoretical themes add another 13.1 percent. Finally when productivity becomes the dependent variable, explained variance increases further given that productivity is more consistent than extremely rare breakthrough events. Indeed with all exploratory variables included, the R^2 raises to 49.6 percent.

Table A.1 – Summary statistics and correlation matrix of dependent and explanatory variables

Variable	N. Obs	Mean	Std. Dev.	Min	Max
nobeldum	3959	0.002	0.039	0	1
top10cite	3959	0.1	0.3	0	1
lncite98	3959	1.761	1.756	0	8.009
lnpub98	3959	0.888	0.822	0	4.477
lnpub97	3959	2.585	1.371	0	7.073
lncite97	3959	4.178	2.058	0	9.692
lnpub97_fl	3959	1.934	1.426	0	6.836
lncite97_fl	3959	3.193	2.3	0	9.533
constraint	3959	0.657	0.315	0	1.932
lncoauthor	3959	2.279	1.15	0	5.855
collabcore	3959	0.821	0.383	0	1
techcore	3959	0.45	0.498	0	1
pubdepth	3959	0.281	0.058	0	1
lexp	3959	2.354	0.883	0	4.174
academic	3959	0.996	0.044	0	1
prestige	3910	5.889	10.683	0	50
prestiged	3910	0.347	0.476	0	1
newcomer	3959	0.036	0.187	0	1
affil1p	3910	0.591	0.492	0	1

	nobeldum	top10cite	lncite98	lnpub98
nobeldum	1			
top10cite	0.1167	1		
lncite98	0.1169	0.6044	1	
lnpub98	0.0168	0.4711	0.815	1

	lnpub97	lncite97	lnpub97_fl	lncite97_fl	constraint	lncoauthor	collabcore
lnpub97	1						
lncite97	0.815	1					
lnpub97_fl	0.939	0.771	1				
lncite97_fl	0.788	0.888	0.861	1			
constraint	-0.21	-0.225	-0.202	-0.216	1		
lncoauthor	0.649	0.568	0.529	0.455	-0.089	1	
collabcore	0.214	0.325	0.151	0.222	-0.361	0.263	1
techcore	-0.219	-0.076	-0.218	-0.104	0.033	0.017	0.063
pubdepth	0.318	0.165	0.302	0.202	-0.069	0.178	-0.005
lexp	0.847	0.724	0.784	0.674	-0.158	0.551	0.252
academic	-0.004	0.02	0.017	0.005	0.001	-0.003	0.008
prestige	-0.058	0.094	-0.044	0.065	-0.014	-0.059	0.066
prestiged	0.1	0.231	0.11	0.206	-0.062	0.016	0.087
newcomer	-0.371	-0.401	-0.268	-0.274	-0.001	-0.392	-0.42
affil1p	0.677	0.599	0.627	0.568	-0.173	0.394	0.166

	techcore	pubdepth	lexp	academic	prestige	prestiged	newcomer	affil1p
lnpub97								
lncite97								
lnpub97_fl								
lncite97_fl								
constraint								
lncoauthor								
collabcore								
techcore	1							
pubdepth	-0.122	1						
lexp	-0.12	0.234	1					
academic	0.011	-0.001	-0.006	1				
prestige	0.084	-0.063	-0.048	-0.027	1			
prestiged	0.024	0	0.108	-0.019	0.755	1		
newcomer	-0.178	0.001	-0.519	0.017	-0.108	-0.143	1	
affil1p	-0.165	0.188	0.698	0.002	-0.056	0.122	-0.235	1

Table A.2 – Predictive models of the Nobel winning paper with rare events logit, top ten percent of citations with logit, number of forward citations of 98 papers and number of 98 papers both with quasi-maximum likelihood Poisson

DV	Relogit Nobel nobeldum b/se	Logit Top10c top10cite b/se	QML impact ncite98 b/se	QML prod npub98 b/se
lnpub97	-1.569**	0.800**	0.316**	1.134**
	(0.32)	(0.12)	(0.07)	(0.04)
lncite97	1.450**	1.153**	0.671**	-0.054**
	(0.27)	(0.08)	(0.04)	(0.02)
lncoauthor	0.536+	-0.094	-0.052	-0.015
	(0.30)	(0.08)	(0.04)	(0.02)
constraint	-4.073*	-0.910**	-0.530**	-0.230**
	(1.86)	(0.30)	(0.15)	(0.08)
collabcore		0.012	0.135	0.084
		(0.29)	(0.12)	(0.07)
techcore	-1.714	-0.036	-0.133	0.014
	(1.57)	(0.14)	(0.10)	(0.04)
pubdepth	12.731**	-0.149	2.075+	0.285
	(4.70)	(1.59)	(1.23)	(0.39)
lexp	-1.788+	-2.015**	-1.136**	-0.962**
	(0.95)	(0.21)	(0.15)	(0.05)
prestige	0.036+	0.014+	0.008+	0.001
	(0.02)	(0.01)	(0.00)	(0.00)
prestiged		-0.057	-0.116	-0.078
		(0.19)	(0.11)	(0.06)
academic		1.734	0.277	0.979**
		(2.18)	(0.74)	(0.38)
newcomer	3.019	2.922**	1.932**	0.321**
	(2.62)	(0.63)	(0.44)	(0.12)
affil1p		0.26	0.315+	0.071
		(0.24)	(0.19)	(0.08)
constant	-7.932*	-6.991**	0.919	-0.921*
	(3.29)	(2.28)	(0.81)	(0.41)
N.Obs	3910	3910	3910	3910
Log-Likelihood		-829.552	-94554.127	-7611.202

+ p<0.10, * p<0.05, ** p<0.01

Appendix B: Interview Question Guide

Open-Ended Questions

Describe your work leading to 1997, in 1998 and after 1998.

Probing Questions on Breakthrough

In the period of 1997-1998 were you and your peers aware that a breakthrough was about to be discovered? Was there excitement due to a potential impactful discovery?

Were scientists trying to solve a specific puzzling mechanism or did they just happen to stumble on the RNAi mechanism by chance while looking for something else?

Were there many teams working towards solving the same problem? Was there racing?

Do you feel like the breakthrough could have been made earlier? Why? What was the missing link that prevented it?

Was the discovery and its results a surprise? In terms of simplicity or complexity of the solution, in terms of who made the discovery?

Before you chose your research direction, how do you evaluate the potential impact of your research? How?

What papers or findings spurred your interest in RNAi research? What works had a decisive influence on your research interests?

What experiments, field or prior breakthroughs do you believe paved the road to the discovery? What inventions (tools), environment fostered the discovery?

Were you aware of the similar co-suppression and quelling results obtained in plants and fungi? / As a plant scientist did you think that co-suppression and quelling would be present in animals?

Probing Questions on the Community

Was there a defined community of RNAi scientists prior to breakthrough?

How would you define the community of RNAi scientists prior to breakthrough? Which subfields of biology came together to form such a community?

How would you characterize this community? Social, open or collective?

How open was the community of scientists working towards solving this discovery? Was there an informal group established that frequently communicated and shared their ideas? Or were results withheld?

What kind of conference/research seminars did you attend at the time, was it phenomenon-based, organism-based or something else?

How do you think about conferences? What role do conferences play in your research?

In your opinion, did the breakthrough come from within the community or from outside?

In your opinion, who were the big contenders in the community to discover the mechanism to RNAi? Why?