Does Science Advance One Funeral at a Time?

Pierre Azoulay
MIT and NBER
Sloan School of Management
100 Main Steet – E62-487
Cambridge, MA 02142
USA

Christian Fons-Rosen
Universitat Pompeu Fabra
Department of Economics
Carrer Ramon Trias Fargas, 25-27
08005 Barcelona
Spain

Joshua S. Graff Zivin
UCSD and NBER
IRPS
9500 Gilman Drive
La Jolla, CA 92093
USA

July 4, 2015

Abstract

Sort of. We study the extent to which eminent scientists shape the vitality of their fields by examining entry rates into the fields of 452 academic life scientists who pass away while at the peak of their scientific abilities. Key to our analyses is a novel way to delineate boundaries around scientific fields by appealing solely to intellectual linkages between scientists and their publications, rather than collaboration or co-citation patterns. Consistent with previous research, the flow of articles by collaborators into affected fields decreases precipitously after the death of a star scientist (relative to control fields). In contrast, we find that the flow of articles by non-collaborators increases by 8% on average. These additional contributions are disproportionately likely to be highly cited. They are also more likely to be authored by scientists who were not previously active in the deceased superstar's field. Overall, the evidence is consistent with the idea that eminent scientists can regulate entry into their field. The mechanism appears to be indirect control: entry in treated fields increases disproportionately when the star had relatively few collaborators in positions of power.

Keywords: economics of science, superstars, invisible college, necromancy.

*Mikka Rokkanen provided additional research assistance. The project would not have been possible without Andrew Stellman's extraordinary programming skills (`www.stellman-greene.com`). We thank Heidi Williams, Scott Stern, and various seminar audiences for useful discussions. The usual disclaimer applies.

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it."

Max Planck
Scientific Autobiography and Other Papers

1 Introduction

Knowledge accumulation—the process by which new research builds upon ideas developed in prior research—has been long understood to be of central importance to scientific progress and economic growth (Mokyr 2002). In deference to Sir Isaac Newton, this cumulative process is often referred to as "standing on the shoulders of giants," but is conceptualized more prosaically as the way in which researchers in one generation learn from and build upon prior research. Yet the literature is largely silent on the mechanisms that shape this slowly evolving process.[1] To borrow terminology from the economic pioneers in the field (Nelson 1962), we know far more about the determinants of the *rate* than that of the *direction* of scientific progress.

Does science evolve according to autonomous laws, or is the direction of science influenced by individuals, incentives, and institutions? While the influence of "external factors" surely matters in the short- and medium-run, philosophers and historians have long argued about the extent to which the pragmatic success of a scientific theory determines how quickly it gains adherents, or its longevity. The epigraph of this paper encapsulates the jaundiced view, attributed to Planck, that the idiosyncratic stances of individual scientists can do much to alter, or at least delay, the course of scientific advance. Yet, the proposition that established scientists are slower than younger ones in accepting novel theories has received little empirical support whenever it has been put to the test (Hull et al. 1978; Gorham 1971; Levin et al. 1995). Moreover, in contrast to technology development where market forces shape the direction of research effort (however imperfectly, cf. Acemoglu [2012]), the choice of a problem-solving approach in basic research is not informed by clear market signals, and thus necessarily depends on a more nuanced system of non-pecuniary incentives (Feynman

[1] This stands in contrast to paradigm-shifting scientific revolutions, which are exceedingly rare but garner far more scholarly attention (e.g., Kuhn (1970), Laudan (1977), and their many detractors). Bramoullé and Saint-Paul (2010) provide an equilibrium model of scientific revolutions with a Kuhnian flavor.

1999). What then guides researchers when choosing between various approaches to study a given problem? And why does the research community coalesce around specific approaches to solving a problem at a particular moment in time?

In this paper, we test Planck's Principle by examining how the death of 452 eminent academic life scientists alter the vitality (measured by publication rates and funding flows) of the subfields in which these scientists actively published in the years immediately preceding their premature passing. Consistent with prior research (Azoulay et al. 2010; Oettl 2012; Jaravel et al. 2015), we find precipitous declines in publication rates in these subfields, relative to control subfields, when we restrict the publication counts to articles authored by collaborators of the stars. Remarkably, however, these declines are more than offset by increased publication rates when we restrict the publication counts to articles authored by non-collaborators. The rest of the manuscript tries to elucidate the mechanisms responsible for this phenomenon.

Our results indicate that these additional contributions by non-collaborators are disproportionately likely to be highly cited, and to represent their authors' first foray into the extinct star's subfield. They also are less likely to cite previous research in the field, and especially less likely to cite the deceased star's work at all. Though not necessarily younger on average, these scientists are also less likely to be part of the scientific elite at the time of the star's death. Interestingly, we do not find evidence that the release of intellectual energy into a field depends on the extent of the star's accomplishments (citations or publications) while alive. It is also implausible that the extinct stars exerted direct control over entry into their fields, since only a vanishingly small number were journal editors or members of NIH study sections. Rather, our results are most consistent with a form of soft, indirect control, whereby potential entrants steer clear of fields in which stars can count on a large number of key collaborators in a position to channel resources (such as editorial goodwill or funding) to insiders. Though stars may have been a source of dynamism while alive, the turnover in leadership enables the injection of fresh ideas into the subfield, thus allowing it to flourish in new directions.

Although it has become a commonplace insight that the process of knowledge accumulation unfolds within an intellectual space, it has proven surprisingly difficult for social scientists to gain empirical traction on this idea (see Borjas and Doran [Forthcoming] for a rare exception). To our knowledge, this is the first paper to examine the dynamics of scientific

evolution using the standard empirical tools of applied microeconomics.[2] We conceptualize the death of eminent scientists as shocks to the structure of the intellectual neighborhoods in which they worked several years prior to their death, and implement a procedure to delineate the boundaries of these neighborhoods in a way that is scalable, transparent, and does not rely on ad hoc human judgment. The construction of our dataset relies heavily on the *PubMed Related Citations Algorithm* [PMRA], which groups scientific articles into subfields based on their intellectual content using very detailed keyword information as well as the relative frequencies of these keywords in the scientific corpus.[3] As such we are able to define circumscribed areas of scientific inquiry that are independent of training, personal relations, or self-proclaimed areas of expertise.

In addition to providing evidence regarding a central question for scholars studying the scientific process, our paper is a departure for the field of the economics of science in that it can attend to the ways in which scientists position themselves simultaneously in an intellectual space as well as a social space, whose boundaries do not overlap. As such, our work can be understood as integrating the traditional concerns of economists—understanding how incentives and institutions influence the rate of knowledge production or diffusion—with those of cognate disciplines such as sociology and philosophy, who have traditionally taken the direction of scientific change as the central problem to be explained.

The rest of the paper proceeds as follows. In the next section, we examine the institutional context and lay out our broad empirical strategy. In section 3, we then turn to data, methods and descriptive statistics. We report the results in section 4. Section 5 concludes by outlining the implications of our findings for future work.

2 Institutional Context and Empirical Design

The setting for our empirical work is the academic life sciences. This sector is an important one to study for several reasons. First, the field has been an enormous source of scientific discovery in the past several decades and continues to play a significant role in the

[2]Considerable work outside of economics has examined the evolution of scientific fields through data visualization techniques (cf. Chavalaris and Cointet (2013) for a recent example). While interesting, this work has been largely descriptive and mostly silent regarding the behavioral mechanisms that might explain the birth, fusion, split, or death of scientific fields.

[3]Unlike in economics, keywords for all publications indexed by *PubMed* (most of the life sciences) are assigned by staff at the National Library of Medicine and are drawn from a controlled vocabulary thesaurus. Thus, concerns about strategic or endogenous keyword choices are minimized in this setting.

health care economy, which accounts for roughly 15% of US GDP. Much biomedical innovation is science-based (Henderson et al. 1999), with the National Institutes of Health (NIH) providing nearly $30 billion in basic science research support in 2014 alone.

Second, the life science research workforce is exceedingly large and specialized. Academic medical centers in the United States employ 150,000 faculty members. Moreover, the discoveries over the past half-decade have greatly expanded the knowledge frontier, necessitating increasing specialization by researchers and a greater role for collaboration (Jones 2009). If knowledge and techniques remain at least partially tacit long after their initial discovery, tightly-knit research teams may be able to effectively control entry into intellectual domains. The size and maturity of this sector, including its extensive variety of narrowly-defined subfields, makes it an ideal candidate for an inquiry into the determinants of the direction of scientific effort in general, and how it is influenced by elite scientists in particular.

Third, the academic research setting also offers the practical benefits of an extensive paper trail of research inputs, outputs, and collaboration histories. On the input side, reliance of researchers on one agency for the majority of their funding raises the possibility that financial gatekeeping by elite scientists could be used to regulate entry into scientific fields. Data on NIH funding at the individual level, as well as membership in "study sections" (the peer-review panels that evaluate the scientific merits of grant applications) will allow us to examine such concerns directly. Most importantly for our study, the principal output of researchers–publications–are all indexed by a controlled vocabulary of keywords managed by the National Library of Medicine. This provides the raw material that allows us to define scientific subfields in a way that is stripped of "social baggage" (the specifics of this process will be described in detail in Section 2.2).

Lastly, while accounts by practicing scientists indicate that collaboration plays a large role in both the creation and diffusion of new ideas (Reese 2004), historians of science have long debated the role of controversies and competition in shaping the direction of scientific progress and the process through which new subfields within the same broad scientific paradigm are born and grow over time (Hull 1989; Morange 1999; Shwed and Bearman 2010). Our study present a unique opportunity to test some of their insights in a way that is more systematic and can yield generalizable insights on the dynamics of field evolution.

3 Empirical Design, Data, and Descriptive Statistics

Below, we provide a detailed description of the process through which the matched scientist/subfield dataset used in the econometric analysis was assembled. We begin by describing the criteria used to select our sample of superstar academics, with a particular focus on "extinction events"; the set of subfields in which these scientists were active prior to their death and the procedure followed to delineate their boundaries. Finally, we discuss the matching procedure implemented to identify control subfields associated with eminent scientists who did not pass away but are otherwise similar to our treatment group.

3.1 Superstar sample

Our basic approach is to rely on the death of "superstar" scientists as a lever to estimate the extent to which the production of knowledge in the fields in which they were active changes after their passing. The study's focus on the scientific elite can be justified both on substantive and pragmatic grounds. The distribution of publications, funding, and citations at the individual level is extremely skewed (Lotka 1926; de Solla Price 1963) and only a tiny minority of scientists contribute through their published research to the advancement of science (Cole and Cole 1972). Stars also leave behind a corpus of work and colleagues with a stake in the preservation of their legacy, making it possible to trace back their careers, from humble beginnings to wide recognition and acclaim.

We began by demarcating a set of 12,935 "elite" life scientists (roughly 5% of the entire relevant labor market) who are so classified if they satisfy at least one of the following criteria for cumulative scientific achievement: (1) highly funded scientists; (2) highly cited scientists; (3) top patenters; and (4) members of the National Academy of Sciences or of the Institute of Medicine.

These four criteria will tend to select seasoned scientists, since they correspond to extraordinary achievement over an entire scientific career. We combine these measures with three others that capture individuals who show great promise at the early and middle stages of their scientific careers, whether or not these episodes of productivity endure for long periods of time: (5) NIH MERIT awardees; (6) Howard Hughes Medical Investigators; and (7) early career prize winners. Appendix I provides additional details regarding these seven metrics of "superstardom."

We trace back these scientists' careers from the time they obtain their first position as independent investigators (typically after a postdoctoral fellowship) until 2006. We do so through a combination of curriculum vitæs, NIH biosketches, *Who's Who* profiles, accolades/obituaries in medical journals, National Academy of Sciences biographical memoirs, and Google searches. For each one of these individuals, we record employment history, degree held, date of degree, gender, and department affiliations.[4]

The 452 scientists who pass away prematurely, and are the particular focus of this paper, constitute a subset of this larger pool of 12,935. Their deaths must intervene between 1975 and 2003 (this allows us to observe at least 3 years' worth of scientific output for every subfield after the death of a superstar scientist). Although we do not impose any age cutoff, the median and mean age at death is 61 with 85% of these scientists having passed away before the age of 70 (we will explore the sensitivity of our results to the superstar's scientist age at death later). We do require evidence, in the form of published articles and/or NIH grants, that these scientists had not entered a pre-retirement phase of their career prior to the time of their death (this is the narrow sense in which we deem their deaths to have occurred prematurely). We painstakingly investigate each extinction event in the sample to determine its cause. This is less difficult that it might seem, since the vast majority of obituaries mention the cause of death explicitly.[5] 229 (51%) of these scientists pass away after a protracted illness, whereas 185 (41%) die suddenly and unexpectedly. We were unable to ascertain the particular circumstances of 37 (8.20%) death events.

Table I provides descriptive statistics for the extinct superstar sample. The median star received his degree in 1957, died at 61 years old and was associated with 4 distinct subfields in the five years leading up to his/her death. On the output side, the stars each received an average of roughly 16.6 million dollars in NIH grants, and published 138 papers that garnered 8,347 citations (as of early 2014).

[4]Though we apply the term of "superstar" to the entire group, there is substantial heterogeneity in intellectual stature within the elite sample (see Table 1).

[5]We exclude from the sample one scientist who took his own life, and a further two for whom suicide could not be ruled out. In ten other instances, the cause of death could not be ascertained from the obituaries and we contacted former collaborators individually to clarify the circumstances of the superstar's passing.

3.2 Delineating Research Fields

The source of the publication data is *PubMed*, an online resource from the National Library of Medicine that provides fast, free, and reliable access to the biomedical research literature.

To delineate the boundaries of the research fields in which each deceased star was active, we develop an approach based on topic similarity as inferred by the overlap in keywords between each articles the star published in the five years prior to his/her death, and the rest of the scientific literature. Specifically, we use the *PubMed Related Citations Algorithm* (PMRA) which relies heavily on Medical Subject Headings (MeSH). MeSH terms constitute a controlled vocabulary maintained by the National Library of Medicine that provides a very fine-grained partition of the intellectual space spanned by the biomedical research literature. Importantly for our purposes, MeSH keywords are assigned to each scientific publication by professional indexers and not by the authors themselves; the assignment is made without reference to the literature cited in the article. We then use the "Related Articles" function in *PubMed* to harvest journal articles that are intellectually proximate to star scientists' own papers.[6]

To fix idea, consider "The transcriptional program of sporulation in budding yeast," an article published in the journal *Science* in 1998 originating from the laboratory of Ira Herskowitz, an eminent UCSF biologist who died in 2003 from pancreatic cancer. As can be seen in Figure I, PMRA returns 72 original journal articles (i.e., excluding reviews, editorials, etc.) for this single source publication. Some of these intellectual neighbors will have appeared before the source to which they are related, whereas others will have only been published after the source. Some will represent the work of collaborators, past or present, of Herskowitz's, whereas others will represent the work of scientists in his field he may never have come in contact with during his life, much less collaborated with. The salient point is that nothing in the process through which these related articles are identified biases us towards (or away from) articles by collaborators, frequent citers of Herskowitz's work, or co-located researchers. Rather, the only determinants of relatedness are to be found in the overlap in MeSH keywords between the source and its potential neighbors.

[6]To facilitate the harvesting of *PubMed*-related records on a large scale, we have developed an open-source software tool that queries *PubMed* and PMRA and stores the retrieved data in a MySQL database. The software is available for download at `http://www.stellman-greene.com/FindRelated/`.

Consider now the second most-related article to Herskowitz's *Science* paper listed in Figure I, "Phosphorylation and maximal activity of *Saccharomyces cerevisiae* meiosis-specific transcription factor Ndt80 is dependent on Ime2." Figure C1 in Appendix C displays the MeSH terms and substances that tag this article along with its source. As a byproduct, PMRA also provides a cardinal dyadic measure of intellectual proximity between each related article and its associated source article. In this particular instance, the relatedness score of "Phosphorylation..." is 94%, whereas the relatedness score for the most distant related article in Figure I, "Catalytic roles of yeast..." is only 62%.

In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications.[7] For each of these publications, we treat the set of publications returned by PMRA as constituting a distinct subfield, and we create a star/field panel dataset by counting the number of related articles in each of these subfields in each year between 1975 and 2006. Of course, these subfields are not necessarily independent in a statistical sense, a fact we will account for in the econometric analysis by clustering all inferences at the level of the star scientist.

3.3 Identification Strategy

A natural starting point to identify the effect of superstar death on entry into scientific subfields is to examine changes in published research output after the superstar passes away, relative to when s/he was still alive, using a subfield fixed effects specification. Since the extinction effect is mechanically correlated with the passage of time, as well as with a subfield's age, our specifications must include life cycle and period effects, as is the norm in studies of scientific productivity (Levin and Stephan 1991). In this framework, the control group that pins down the counterfactual age and calendar time effects for the subfields that currently experience the death of a superstar consists of subfields whose associated superstar died in earlier periods, or will die in future periods. If the death of a superstar only represented a one-time shift in the level of entry into the relevant subfields, this would not be problematic. Let us assume instead that extinction events affect trends, and not just levels, of scientific activity in a subfield. Relying on a single level of difference for identification would be tantamount to assuming that age and calendar time effects for the treated subfields can

[7] A robust social norm in the life sciences systematically assigns last authorship to the principal investigator, first authorship to the junior author who was responsible for the actual conduct of the investigation, and apportions the remaining credit to authors in the middle of the authorship list, generally as a decreasing function of the distance from the extremities of the list.

completely filter out the effect of the star's passing from that of the mere passage of time. This is a burden that these time and age effects should not be expected to bear in the presence of idiosyncratic life cycle patterns at the level of a subfield (e.g., with their productive potential first increasing over time, eventually peaking, and thereafter slowly declining).

To mitigate this threat to identification, our preferred empirical strategy relies on the selection of a matched scientist/subfield for each treated scientist/subfield. These control observations are culled from the universe of of superstars who do not die (see Section 2.1 and Appendix D). Combining the treated and control samples enables us to estimate the effect of superstar extinction in a difference-in-differences framework. Figure II illustrates the procedure used to identify control subfields in the particular case of the Herskowitz's publication highlighted above.

We begin by looking at all the articles that appeared in the same journals and in the same year as the treated source articles. From this set of articles, we keep only those that have one of the still-living superstar in last authorship position. Then, using a "coarsened exact matching" procedure detailed in Appendix D, the control source articles are selected such that (1) the number of authors in the treated and control are approximately similar; (2) the age of the treated and control superstars differ by no more than five years; and (3) the number of citations received by the treated and source article are similar. For the Herskowitz/"sporulation in budding yeast" pair, we can select 10 control articles in this way. All of these controls were also published in *Science* in 1998, and have between five and seven authors. One of these controls is "Hepatitis C Viral Dynamics in Vivo...," whose last author is Alan Perelson, a biophysicist at Los Alamos National Lab. Perelson and Herskowitz obtained their PhD only a year apart. The two papers had received 514 and 344 citations respectively by the end 2003. Though this is a large difference, this places both well above the 99^{th} percentile of the citation distribution for 5-year old articles published in 1998.

One potential concern with the addition of this "explicit" control group is that control subfields could be affected by the treatment of interest. What if, for instance, a control source article happens to be related (in a PMRA sense) with the treated source? Because the subfields identified by PMRA are narrow, this turns out to be an infrequent occurrence. Nonetheless, we remove all such instances from the data. We then find all the intellectual neighbors for these control source articles using PMRA, and a control subfield is defined by the set of related articles returned by PMRA, in a manner that is exactly symmetric to the

procedure used to delineate treated subfields. When these related articles are parsed below to distinguish between those published by collaborators vs. non-collaborators of the star, or between those by intellectual outsiders vs. insiders, treated and control observations will always be defined with perfect symmetry.

3.4 Descriptive Statistics

The procedure described above yields a total of 34,071 distinct subfields; 3,071 subfields correspond to one of the 452 extinct scientists, whereas 31,102 subfields correspond to one of 5,797 still-living scientists. Table II provides descriptive statistics for control and treated subfields in the baseline year, i.e., the year of death for the extinct scientist.[8]

Covariate balance. In the list of variables displayed in Table II, it is important to remember that a number of covariates are balanced between treated and control subfields solely by virtue of the coarsened exact matching procedure—for instance, (star) investigator year of degree, or the source article number of authors, or the source article number of citations at baseline.

However, there is nothing mechanical to explain the balance between treated and control subsamples with respect to the stock of our main outcome variable: the number of articles in the field. Figure III compares the corresponding distribution and also shows a great deal of overlap between the two histograms. Of course, balance in the levels of the outcome variable is not technically required for the validity of the empirical exercise.[9] Yet, given the ad hoc nature of the procedure used to identify control subfields, this degree of balance is quite reassuring.

Another happy byproduct of our matching procedure is that treated and control scientists also appear quite similar in the extent of their eminence at the time of (counterfactual) death, whether such eminence is measured through NIH funding, the number of articles published, or the number of citations these articles received.

[8] We can assign a counterfactual year of death for each control subfield, since each control subfield is associated with a particular treated subfield through the matching procedure described above.

[9] What is required is that the trends in publication activity be comparable between treated and control subfields up until the death of the treated scientist. We verify that this is the case below.

Collaborators vs. non-collaborators. One critical aspect of the empirical analysis is to distinguish between collaborators and non-collaborators of the star when measuring publishing activity in a subfield. It is therefore crucial to describe how this distinction can be made in our data. Information about the superstars' colleagues stems from the Faculty Roster of the Association of American Medical Colleges, to which we secured licensed access for the years 1975 through 2006, and which we augmented using NIH grantee information (cf. Azoulay et al. [2010] for more details).

An important implication of our reliance on these source of data is that we can only identify authors who are faculty members in U.S. medical schools, or recipient of NIH funding. In particular, we cannot systematically identify trainees, scientists working for industrial firms, or scientists employed in foreign academic institutions. The great benefit of using these data, however, is that they ensure we know quite a bit about the individuals we are able to identify: their (career) age, type of degree awarded, place of employment, gender, and research output, whether measured by publications or NIH grants.

To identify authors, we match the authorship roster of each related article in one of our subfields with the AAMC roster.[10] We tag as a collaborator any author who appeared as an author on a publication prior to the death with the star associated with the subfield. Each related article is therefore assigned to one of two mutually-exclusive bins: the "collaborator" bin comprises the set of publications with at least one identified author who coauthored with the star prior to the year of death (or counterfactual death); the "non-collaborator" bin comprises the set of publications with no identified author who coauthored with the star prior to the year of death (or counterfactual death). As can be seen in Table II, roughly 15% of the publication activity at baseline can be accounted for by collaborators. Moreover, this proportion is very similar for control and treated subfields.

4 Results

The exposition of the econometric results proceeds in stages. After a brief review of methodological issues, we provide results that pertain to the main effect of superstar exposure on subfield growth, measured by publication rates and funding flows. Second, we attempt to

[10] We limit ourselves to authors with relatively infrequent names. Though we will surely result in some measurement error, there is no reason to suspect that this will impact treated and control subfields in a differential way.

elucidate the mechanism (or set of mechanisms) at work to explain our most robust finding, that of relative subfield growth in the wake of star extinction, a growth entirely accounted for by contributions from non-collaborators. We do so by examining the characteristics of the articles published by non-collaborators, before turning to the characteristics of their authors. We also explore heterogeneity in the treatment effect through the interaction of the post-death indicator variable with various attributes of the stars.

4.1 Econometric Considerations

Our estimating equation relates publication or funding activity in subfield i in year t to the treatment effect of losing superstar j:

$$E\left[y_{it}|X_{ijt}\right] = exp\left[\beta_0 + \beta_1 AFTER\_DEATH_{jt} + f(AGE_{it}) + \delta_t + \gamma_{ij}\right] \quad (1)$$

where y is a measure of activity, $AFTER\_DEATH$ denotes an indicator variable that switches to one in the year during which superstar j passes away, $f(AGE_{it})$ corresponds to a flexible function of the field's age, the δ_t's stand for a full set of calendar year indicator variables, and the γ_{ij}'s correspond to subfield/star fixed effects, consistent with our approach to analyze *changes* in activity within i following the passing of superstar j.

The subfield fixed effects control for many time-invariant characteristics that could influence research activity, such as the need for capital equipment or the extent of disease burden (e.g., for clinical fields). A pregnant metaphor for the growth of scientific knowledge has been that of biological evolution (Hull 1989; Chavalarias and Cointet 2013): a field is born when new concepts are introduced, resulting in an accelerating production of "offsprings" (articles), until the underlying scientific community loses its thematic coherence, ushering in an era of decline (or alternatively, splitting or merging events). To flexibly account for such life cycle effects, we include subfield age indicator variables, where subfield age is computed as the number of years since the year of publication for the underlying article.[11] The calendar year effects filter out the effects of the general expansion of the scientific enterprise as measured by the number of journals and articles published each year.[12]

[11] An alternative way to measure subfield age is to date its birth year as the year during which the first related article was published. Though our main results are robust to this alternative parametrization, this is not a desirable way to proceed since it will fail to distinguish subfields that are genuinely long-established from fields that are more recent but happen to have an ancient precursor that PMRA is able to recognize.

[12] It is not possible to separately identify calendar year effects from age effects in the "within subfield" dimension of a panel in a completely flexible fashion, because one cannot observe two subfields at the same point in time that have the same age but were born in different years (Hall et al. 2007).

Estimation. The dependent variables of interest, including publication counts and NIH grants awarded, are skewed and non-negative. For example, 31.40% of the subfield/year observations in the data correspond to years of no publication activity; the figure climbs to 56.70% if one focuses on the count of NIH grants awarded. Following a long-standing tradition in the study of scientific and technical change, we present conditional quasi-maximum likelihood estimates based on the fixed-effect Poisson model developed by Hausman et al. (1984). Because the Poisson model is in the linear exponential family, the coefficient estimates remain consistent as long as the mean of the dependent variable is correctly specified (Gouriéroux et al. 1984).

Inference. QML (i.e., "robust") standard errors are consistent even if the underlying data generating process is not Poisson. In fact the Hausman et al. estimator can be used for any non-negative dependent variables, whether integer or continuous (Santos Silva and Tenreyro 2006), as long as the variance/covariance matrix is computed using the outer product of the gradient vector (and therefore does not rely on the Poisson variance assumption). Further, QML standard errors are robust to arbitrary patterns of serial correlation (Wooldridge 1997), and hence immune to the issues highlighted by Bertrand et al. (2004) concerning inference in DD estimation. We cluster the standard errors around superstar scientists in the results presented below.

Dependent Variables. Our primary outcome variable is publication activity in a subfield. However, we go beyond this raw measure by assigning the related articles that together constitute the subfield into a variety of bins. For instance, we can decompose publication activity in the subfield into two mutually exclusive subfields: articles that appear in prestigious journals (Journal Impact Factor [JIF] higher than two) and those that appear in less prestigious journals (JIF lower than two); or articles with a superstar on the authorship roster vs. articles without a superstar; etc. Articles in each bin can then be counted and aggregated up to the subfield/year level.

Capturing funding flows at the field level is slightly more involved. *PubMed* systematically records NIH grant acknowledgements using grant numbers, but without referencing the particular grant cycle to which the publication should be credited. To address this issue, we adopt the following rule: for each related publication, we identify the closest preceding year in a three-year window during which funding was awarded through either a new award or a

competitive renewal; we then sum all the awards in the grant year that ultimately generate publications in the focal subfield.

4.2 Main effect of superstar extinction

Table III and Figure IV present our core results. Overall, we find that publication activity increases slightly following the death of a star scientist who was an active contributor to it, but the magnitude of the effect is not large (about 2%) and imprecisely estimated (column 1). Yet, this result conceals a striking pattern that we uncover when we distinguish between publication by collaborators and non-collaborators. The decline in publication activity accounted for by previous collaborators of the star is massive, on the order of 40% (column 2). This evidence is consistent with our previous findings, which showed that coauthors of superstar scientists who die suffer a drop in output, particularly if their non-collaborative work exhibited strong keyword overlap with the star, i.e., if they were intellectually connected in addition to being coauthors (Azoulay et al. 2010, Table VI, column 2).

A limitation of the previous work focusing on the fate of collaborators after the loss of an eminent scientist always lied in the failure to distinguish between social and intellectual channels of influence, since every treated scientist was by definition a collaborator, even if merely a casual one. In this study, we can relax this constraint, and when we do, we find that publication activity by non-collaborators in the subfield increases by a statistically significant 8.00% (column 3).

We also explore the dynamics of the effects uncovered in Table III. We do so by estimating a specification in which the treatment effect is interacted with a set of indicator variables corresponding to a particular year relative to the superstar's death, and then graphing the effects and the 95% confidence interval around them (Panels A, B, and C of Figure IV correspond to columns 1, 2, and 3 in Table III). Two features of the figure are worthy of note. First, the dynamics amplify the previous results, in the sense that we see the effects increasing (in absolute value) monotonically over time—there is no indication that the effects we estimated in Table III are merely transitory. Five years after a star's death, the increase in publication activity by non-collaborators is large enough in magnitude to fully offset the

decline in activity by collaborators. Second, there is no discernible evidence of an effect in the years leading up to the death, a finding that validates ex post our identification strategy.[13]

The last three columns of Table III focus on funding flows from the National Institutes of Health (NIH) rather than publication flows. More precisely, the outcome variables in columns 4, 5, and 6 is the number of distinct NIH awards that acknowledge a publication in the subfield in the three-year window before the year of publication for the related article (counting grant amounts, as opposed to the number of grants, yields similar results). Note that the sample size is a bit smaller, since the conditional fixed effects Poisson specifications drop from the estimation sample all subfields with no recorded history of NIH grant receipt. The patterns are very similar to those obtained in the case of publication activity, both in terms of magnitudes and in terms of statistical significance.[14]

4.3 Understanding extinction-induced subfield growth

In the remainder of the manuscript, we seek to understand the mechanisms that might explain the novel empirical regularity we uncovered: that of relative growth for subfields following the death of their superstar anchor, a phenomenon entirely accounted for by research activity undertaken by scientists who never collaborated with the star while alive. As a consequence, all the results below pertain to entry into the field by non-collaborators; any article with even one author who collaborated with the star in the past, present, or future is excluded from the count of articles that constitute the dependent variable.

Article Characteristics. What characterizes the additional contributions that together lead to increased activity in a subfield following star extinction? Are these in fact important contributions to the subfield? Do they focus on core issues, or should they be understood as taking the intellectual domain in a novel direction? Tables IV and V explore these issues. In Table IV, we parse every related article that constitute the subfields in our data to assign them into one of six mutually exclusive bins, based on long-run citation impact: articles that fall in the bottom quartile of the citation distribution; in the second quartile; in the third quartile; articles that fall above the 75^{th} percentile, but below the 95^{th} percentile; articles

[13]This finding is reassuring as it suggests that death events are plausibly exogenous to the course of knowledge growth and decline within a subfield. The case for exogeneity is stronger in the case of sudden death than in the case of anticipated death, a distinction that we will examine in more detail below.

[14]The event study graphs corresponding to the dynamics of funding flows are available from the authors, but also show close similarity to those displayed in Figure IV.

that fall above the 95^{th} percentile, but below the 99^{th} percentile; articles that fall above the 99^{th} percentile of the citation distribution.[15]

Panel A of Table IV produces a battery of estimates corresponding to each of these six bins in columns 2 through 7 (column 1 simply replicates the effect for all papers, regardless of impact, that was previously displayed in Table III, column 3). A startling result is that the magnitude of the treatment effect increases sharply as we focus on the rate of contributions with higher impact. In contrast, the number of lower-impact articles contributed by non-collaborators contracts slightly, though the effect is not precisely estimated.

Panels B and C break down these results further by examining separately the growth of subfields by cause of death (anticipated vs. sudden). As mentioned earlier, the case for exogeneity is stronger in the case of sudden death, since when the death is anticipated, it would be theoretically possible for the star to engage in "intellectual estate planning," whereby particular scientists (presumably close collaborators) are anointed as representing the next generation of leaders in the subfield. The results in column 1 imply that there is an important difference between the two type of events—subfield growth is observed mostly when the death of the star was anticipated. Decomposing this effect across the quantile bins as above reveals that these differences can be accounted for by shifts in activity for low-impact contributions. In the right tail of the distribution, there is very little evidence that the manner of superstar death matters at all for the fate of their subfields. In both cases, non-collaborators increase their contribution sharply—on the order of 40%. Because of this convergence in the upper tail, the remainder of the manuscript will lump together anticipated and unanticipated events.[16]

Table V parses the related articles in each subfield to ascertain whether contributions by non-collaborators constitute a genuine change in intellectual direction. Panel A distinguishes between contributions that are proximate in intellectual space to the source article vs. those that are slightly more distant (though still part of the subfield as construed by PMRA).

[15]Note that when we are referring to the citation distribution, we mean the vintage-specific citation distribution for the universe of articles simultaneously indexed by *PubMed* and the Web of Science. For example, the article by Sopko et al. highlighted on Figure C1 (in Appendix C) received 39 citations from other articles in *PubMed* by 2014. This puts this article above the 76^{th} percentile of the citation distribution for articles published in 2002.

[16]The most salient results reported below continue to hold when analyzed separately by cause of death. However, we gain statistical power from pooling these observations, and some empirical patterns would be estimated less precisely if we chose to focus solely on observations corresponding to subfields for which the star died suddenly and unexpectedly.

Because we have at our disposal both a cardinal and an ordinal measure of intellectual proximity, we present four different estimates. In both cases, the magnitude of the treatment effect pertaining to publication activity by proximate articles is approximately twice as large as the magnitude corresponding to more distant articles. These differences, however, are not themselves statistically significant at conventional levels. But we can at least rule out the conjecture that non-collaborators enter the field from the periphery. Their contributions seem to lie smack-dab in the middle of the subfield as it existed when the star was still alive.

Panel B sheds light on the intellectual direction of the field, by examining the cited references contained in each related article. The first two columns separate related articles in two groups. The first contains only publications that cite at least some work which belongs to the subfield identified by PMRA for the corresponding source. The second contains publications that cite exclusively out of the PMRA subfield. Only articles in the second group appear to experience growth in the post-extinction era. The next two columns proceed similarly, except that the list of references is now parsed to highlight the presence of articles authored by the star, as opposed to all other authors. We find that subfield growth can be mostly accounted for by articles from non-collaborators who do not build on the work of the star. Finally, we investigate the vintage of the references cited by related articles. The last two columns in Panel B indicate that the new contributions are more likely to build on science of a more recent vintage.

Taken together, the results in Panels A and B of Table V paint a nuanced picture of directional change in the wake of superstar extinction. The new contributions do not represent a radical departure from the subfield's traditional questions—their MeSH keywords overlap with those of the source article even more than is typical for the "average" article in the subfield. At the same time, the citation evidence makes it clear that these additional contributions often draw from more recent and novel sources of knowledge for inspiration.

Related Author Characteristics. The next step of the analysis is to investigate the type of scientists who publish the articles that account for subfield growth in the wake of a star's death. Table VI reports these results. Perhaps the simplest author characteristic is age. For each related article in the subfield, we match the authorship roster to the AAMC Faculty Roster. Then, we compute the mean career age over matched authors for each related article. Since the median career age for matched authors turns out to be 16, we assign each article to one of two bins, the first comprising all related articles with an "older" authorship team

(mean author career age greater than 16), the second comprising all related articles with a "younger" authorship team (mean author career age less than or equal to 16). We then compute publication activity separately for these two groups by aggregating these data up to the subfield/year level of analysis. As can be observed in the first two columns of Table VI, there really is not any difference in the magnitude of the extinction effect across these two groups.

The second step is to distinguish between the related articles with at least one eminent author from related articles for which none of the authors is particularly famous at the time of its publication. To do this, we use two distinct measures of eminence. The first is whether a matched author belong to our sample of 12,935 stars. The second is whether a matched author belongs to an even more elite set comprising Nobel prize winners, Howard Hughes Medical Investigators, and members of the National Academy of Sciences. In the final four columns to Table VI, we find that articles published by non-elite members of the profession appear to account for much of the relative growth for treated subfields. This is consistent with the idea that elite scientists face weaker incentives to deviate from their existing research trajectory, relative to less-established scientists.

Finally, we probe the standing of the non-collaborators in the subfield. One possibility is that they are competitors of the star, with much of their publication activity in the subfield when the star was alive. Another possibility is that they are recent entrants into the subfield—not social outsiders but intellectual outsiders. To distinguish these different types of authors empirically, we create a metric of intellectual proximity for each matched author, by computing the fraction of their publications that belongs to the star's subfields up to the year before the publication of each related article. Whenever we match more than one author on a single related article, we assign to that article the maximum proximity score. A full 50% of the related articles turn out to have authors with exactly zero intellectual overlap with the star's subfield. In addition to the bottom two quartiles, we create 10 bins for every five percentiles above the median (50^{th} to 55^{th} percentile, 55^{th} to 60^{th} percentile,..., 95^{th} to 99^{th} percentile), as well as top percentile bin. We then compute the corresponding measures of subfield activity by aggregating the data up to the subfield/year level. This time, we opt to present the results graphically in Figure V. Each dot corresponds to the magnitude of the treatment effect in a separate regression with the outcome variable being the number of articles in each subfield that belong to the corresponding bins.

A striking pattern emerges. The authors driving the growth in publication activity following a star's death are largely outsiders. They do not appear to have been substantially active in the subfield when the star was alive. To borrow a term from industrial organization, they are new entrants into these subfields, though the evidence presented above also shows that they are not especially likely to be younger scientists overall.

Star Characteristics. The evidence so far points to fields of deceased stars enjoying bursts of activity after the extinction event. Does the magnitude of these bursts depend on characteristics of the star? Table VII reports the relevant results. In these analyses, we split the sample at the median of a series of covariates defined at the level of the star, or at the level of the star/subfield pair (at baseline). A natural starting point is to investigate the relationship between the magnitude of the treatment effect and the accomplishments of the star while alive. We rank superstars according to career citations (normalized by the career age of the star at the time of death).[17] The first two columns of Table VII split the sample at the median of the career citation count for each star. We do find a slightly more pronounced response for the subfields of stars who shined relatively less bright while alive, but the difference between the two coefficient estimates is small and statistically insignificant.

The next two columns investigate the sensitivity of our results to the age of the star at death, using 61 years as a cutoff (a choice which splits the sample in two approximately equal halves). Recall that all star scientists (whether associated with treated or control subfields) meet some threshold level of activity to be included in the sample. Yet, one would expect less of an effect on growth or decline in the case of stars who find themselves in the twilight of their career when they pass away. Indeed, we do find that the treatment effect is much more pronounced for subfields who experience the death of a relatively young superstar. The effect in the subsample corresponding to subfields who experience the death of stars aged 61 and above is statistically indistinguishable from zero.

Next, we examine the influence of the star's importance for the field, computed as the fraction of papers in the subfield that have the star as an author. Perhaps unsurprisingly, we find that subfield growth is more pronounced when the star's passing creates a larger void for non-collaborators to fill. The final two columns of Table VII look at variation in the star's commitment to the field—one might expect that the passing of stars whose main

[17] The same analysis was also conducted using career publications as the metric of achievement, with very similar results.

research interests lay outside the focal subfield has less of an influence, relative to stars with a singular focus on the subfield. To test this hypothesis, we look at the star's entire corpus of work, and ask what fraction of it actually falls in the focal subfield. We fail to detect any meaningful difference along this dimension.

Direct and Indirect Control. We have found that the fields in which the star was active expand following his/her demise, with outsiders, rather than competitors, driving much of the increase in activity. Together, the results suggest that stars might be able to regulate entry into their field while alive; their premature death leaves a void that new entrants into the intellectual domain fill.

While this story makes intuitive sense, it is also speculative. What are the mechanisms that enable these eminent scientists to titrate the number and type of scientists who can enter into—and thrive within—the fields in which they are active? One possibility is that stars can exert control in a direct way through the control of key resources, such as funding or editorial goodwill (Li 2014; Brogaard et al. 2014). This possibility strikes us as highly implausible. In the five year window before the death, only three of our stars (out of 452) were sitting on study sections, the funding panels that evaluate the scientific merits of NIH grant applications. Another three were journal editors in the same time window. This handful of individuals does not drive the robust effects we have uncovered. Alternatively, one might consider the amount of NIH funding received by a star as a reasonable proxy for control over resources. In the first two columns of Table VIII, Panel A, we split our sample into two, the first half corresponding to the subfields of comparatively less well-funded stars, the second corresponding to the subfields of comparatively more well-funded stars.[18] We do find that the increase in activity is more pronounced when the star was relatively less-lavishly funded, but the difference between the two coefficient estimates is not itself statistically significant.

Stars also have the ability to regulate entry into their fields indirectly, through the influence they exert on their collaborators. But these stars tend to have a very large number of collaborators; which among those collaborators could be instrumental in shaping the intellectual direction of the field? The next two columns of Table VIII, Panel A investigate the role of trainees. To identify trainees, we focus on the subset of coauthors who occupy

[18]Note that the sample is slightly smaller for the funding split, since we need to drop from the sample the subfields corresponding to superstars who are NIH intramural scientists, and as such not eligible for extramural NIH funding.

the first author position in articles where the star occupies the last position; with the added stipulation that the coauthored publication appears in a window of ± three years around the year in which the collaborator's highest degree was received. We conjecture that the fields of stars who produced many intellectual "offsprings" might be less welcoming for outsiders, relative to the fields of stars who did not train as many graduate students or postdoctoral fellows. This is indeed what we find, but the result is merely suggestive—the difference between the two coefficients is not statistically different from zero.

To probe the possibility of indirect control further, we delineate two categories of "important" collaborators. The first comprises those individuals who coauthor frequently with the star: five coauthorships or more at the time of the star's death (this corresponds to the top decile of collaborators when ranked by total number of coauthorships). The second uses "extreme" authorship positions, focusing on collaborators who were ever first author when the star was in last position, or last author when the star was in first position. Using information regarding the composition of NIH funding panels, we then tabulate, for each star, the number of important collaborators who were members of at least one of these committees in the five years preceding the death of the star.

We could proceed in a similar fashion using the composition of editorial boards. Unfortunately, these data are not easily available for the set of *PubMed*-indexed journals and the thirty-year time period covered by our sample. We develop a proxy for editorial position based on the number of editorials or comments written by every collaborator of the star.[19] We then sum the number of editorials written by important coauthors in the five years before the extinction event. Together, the editorial and study section information allow us to distinguish between the stars whose important coauthors were in a position to channel resources towards preferred individuals or intellectual approaches, versus those stars whose important coauthors had no such power.

Panel B of Table VIII presents the evidence on the role of indirect control. The eight specifications paint a unified picture—subfield expansion is the rule, but is much more pronounced when indirect control was a possibility. The differences between estimates in each

[19]We investigated the validity of this proxy as follows. In the sample of extinct superstars, every individual with five editorials or more was an editor. In a random sample of 50 superstars with no editorials published, only one was an editor (for a field journal). Finally, among the sixteen superstars who wrote between one and four editorials over their career, we found two whose CV indicate they were in fact editors for a key journal in their field. We conclude that their appears to be a meaningful correlation between the number of editorials written and the propensity to be an editor.

pair of columns is large, and significantly different from zero at the 5% level of significance in one-tailed tests. Indirect control therefore appears to be a mechanism through which superstars can exert influence on the evolution of their fields, even from beyond the grave. The evidence points to the possibility that these important coauthors, in their effort to keep the star's intellectual flame alive, erect barriers to entry into those fields that prevent its rejuvenation by outsiders.

One limitation of this analysis is that we do not have a crisp story to explain why outsiders refrained from entering the field in those instances where indirect gatekeeping was less likely to be effective while the star was alive. Conversely, we do not have a clear explanation as to what prevents the fabric of this network from fraying after the star's passing, thus lowering the intellectual barriers to entry. A possibility is that the presence of a towering figurehead dissuades intellectual outsiders to engage with the field. Upon the disappearance of this figurehead, outsiders may well attempt to enter these fields, but their efforts might only be successful when indirect control is less effective.

5 Conclusion

In this paper, we exploit the applied economist's toolkit, together with a novel approach to delineate the boundaries of scientific fields, to explore the effect that the passing of an eminent scientist on the dynamics of growth—or decline—for the fields in which s/he was active while alive. Consistent with earlier work (Azoulay et al. 2010), we find that the death of an elite scientist has a negative and seemingly permanent impact on the productivity of their coauthors. In contrast, the productivity of non-collaborators within the same fields appears to increase, at a rate that more than offsets the decline experienced by collaborators. Our rich data on individual researchers and the nature of their scholarship allows us provide a deeper understanding of this dynamic.

While coauthors suffer after the passing of a superstar, it is not simply the case that star scientists in a competing lab assume the leadership mantle. Rather, the boost comes largely from outsiders who appear to tackle the mainstream questions within the field but by leveraging newer ideas that arise in other domains. This intellectual arbitrage is quite successful—the new articles represent substantial contributions, at least as measured by long-run citation impact. Together, these results paint a picture of scientific fields as scholarly

guilds to which elite scientists can regulate access, providing them with outsized opportunities to shape the direction of scientific advance in that space.

We also provide evidence regarding the mechanisms that enable the regulation of entry. Rather than active and direct gatekeeping by the star, our analyses suggest a kind of soft power exercised through a tightly-knit "invisible college" of collaborators able to channel resources to insiders through their position on editorial board and funding committees (de Solla Price and Beaver 1966; Crane 1972). The loss of an elite scientist central to the field appears to signal to those on the outside that the cost/benefit calculations on the avant-garde ideas they might bring to the table has changed, thus encouraging them to engage. This phenomenon is especially pronounced when the star's network of collaborators is insufficiently robust to stave off threats from intellectual outsiders.

In the end, our results lend credence to Planck's infamous quip that provides the title for this manuscript. Yet its implications for social welfare are ambiguous. While we can document that eminent scientists restrict the entry of new ideas and scholars into a field, gatekeeping activities could have beneficial properties when the field is in its inception; it might allow cumulative progress through shared assumptions and methodologies, and the ability to control the intellectual evolution of a scientific domain might, in itself, be a prize that spurs much *ex ante* risk taking. Because our empirical exercise cannot shed light on these countervailing tendencies, we must remain guarded in drawing policy conclusions from our results. Yet, the fact that the presence of a tutelar figurehead can freeze patterns of participation into a scientific field increases the appeal of policies that bolster access to less established or less well-connected investigators. Example of such policies include caps on the amount of funding a single laboratory is eligible to receive, "bonus points" for first-time investigators in funding programs, emeritus awards to induce senior scientists to wind down their laboratory activities, and double-blind refereeing policies (Kaiser 2011, Deng 2015).

Our work leaves many questions unanswered. What is the fate of the fields that these new entrants departed? Do they decay, or instead "merge" with those whose star departed prematurely? Given a finite supply of scientists and the adjustment costs involved in switching scientific focus, one would expect some other field to contract on the margin in the wake of superstar extinction. Is this marginal field more novel, or already established? We are pursuing these questions in ongoing work.

References

Acemoglu, Daron. 2012. "Diversity and Technological Progress." In Josh Lerner, and Scott Stern (Eds.), *The Rate & Direction of Inventive Activity Revisited*, pp. 319-356. Chicago, IL: University of Chicago Press.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* **125**(2): 549-589.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* **119**(1): 249-275.

Bramoullé, Yann, and Gilles Saint-Paul. 2010. "Research Cycles." *Journal of Economic Theory* **145**(5): 1890-1920.

Brogaard, Jonathan, Joseph Engelberg, and Christopher Parsons. 2014. "Network position and productivity: Evidence from journal editor rotations." *Journal of Financial Economics* **111**(1): 251-270.

Chavalarias, David, and Jean-Philippe Cointet. 2013. "Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields." *PLoS one* **8**(2): e54847.

Cole, Jonathan R., and Stephen Cole. 1972. "The Ortega Hypothesis." *Science* **178**(4059): 368-375.

Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities.* Chicago, IL: University of Chicago Press.

de Solla Price, Derek J. 1963. *Little Science, Big Science*. New York: Columbia University Press.

de Solla Price, Derek J., and Donald D. Beaver, "Collaboration in an Invisible College." *American Psychologist*, 21 (1966), 1011-1018.

Deng, Boar. 2015. "NIH Ponders Emeritus Grants." Nature **518**(7538): 146-147.

Doran, Kirk, and George Borjas. 2015. "Which Peers Matter? The Relative Impacts of Collaborators, Colleagues, and Competitors." *Review of Economics and Statistics,* Forthcoming.

Feynman, Richard P. 1999. *The Pleasure of Finding Things Out*. New York: Basic Books.

Gorham, Geoffrey. 1991. "Planck's Principle And Jeans's Conversion." *Studies in History & Philosophy of Science* **22**(3): 471-497.

Gouriéroux, Christian, Alain Montfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica* **53**(3): 701-720.

Hall, Bronwyn H., Jacques Mairesse, and Laure Turner. 2007. "Identifying Age, Cohort and Period Effects in Scientific Research Productivity: Discussion and Illustration Using Simulated and Actual Data on French Physicists." *Economics of Innovation and New Technology* **16**(2): 159-177.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica* **52**(4): 909-938.

Henderson, Rebecca, Luigi Orsenigo, and Gary P. Pisano. 1999. "The Pharmaceutical Industry and the Revolution in Molecular Biology: Interactions Among Scientific, Institutional, and Organizational Change." In David C. Mowery, and Richard R. Nelson (Eds.), *Sources of Industrial Leadership*, pp. 267-311. New York: Cambridge University Press.

Hull, David L. 1988. *Science as a Process*. Chicago, IL: University of Chicago Press.

Hull, David L., Peter D. Tessner, and Diamond. Arthur M. 1978. "Planck's Principle." *Science* **202**(4369): 717-723.

Jaravel, Xavier, Neviana Petkova, and Alex Bell. 2015. "Team-Specific Capital and Innovation." Working paper, Harvard University.

Jones, Benjamin 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.

Kaiser, Jocelyn. "Darwinism vs. Social Engineering at NIH." *Science* **334**(6057): 753-754.

Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Laudan, Larry. 1998. *Progress and its Problems: Towards a Theory of Scientific Growth*. Berkeley, CA: University of California Press.

Levin, Sharon G., and Paula E. Stephan. 1991. "Research Productivity over the Life Cycle: Evidence for Academic Scientists." *American Economic Review* **81**(1): 114-32.

Levin, Sharon G., Paula E. Stephan, and Mary Beth Walker. 1995. "Planck's Principle Revisited: A Note." *Social Studies of Science* **25**(2): 275-283.

Li, Danielle. 2015. "Expertise vs. Bias in Evaluation: Evidence from the NIH." Working Paper, Harvard University.

Lotka, Alfred J. 1926. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* **16**(12): 317-323.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.

Morange, Michel. 1998. *A History of Molecular Biology*. Cambridge, MA: Harvard University Press.

Nelson, Richard R. (ed.). 1962. *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton University Press.

Oettl, Alexander. 2012. "Reconceptualizing Stars: Scientist Helpfulness and Peer Performance." *Management Science* **58**(6): 1122-1140.

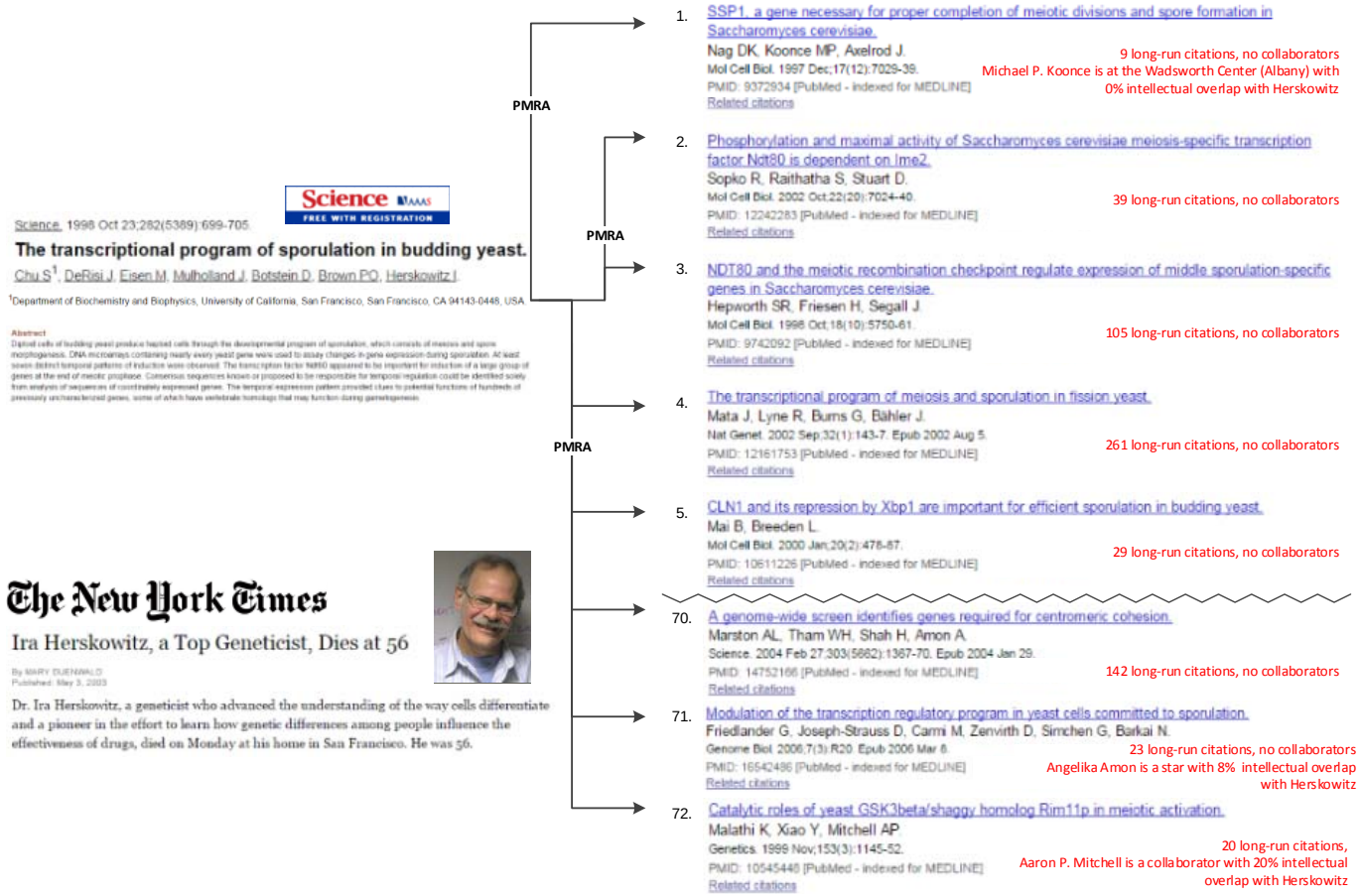
Reese, Thomas S. 2004. "My Collaboration with John Heuser." *European Journal of Cell Biology* **83**(6): 243-244.

Santos Silva, J.M.C., and Silvanna Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* **88**(4): 641-658.

Shwed, Uri, and Peter S. Bearman. 2010. "The Temporal Structure of Scientific Consensus Formation." *American Sociological Review* **75**(6): 817-840.

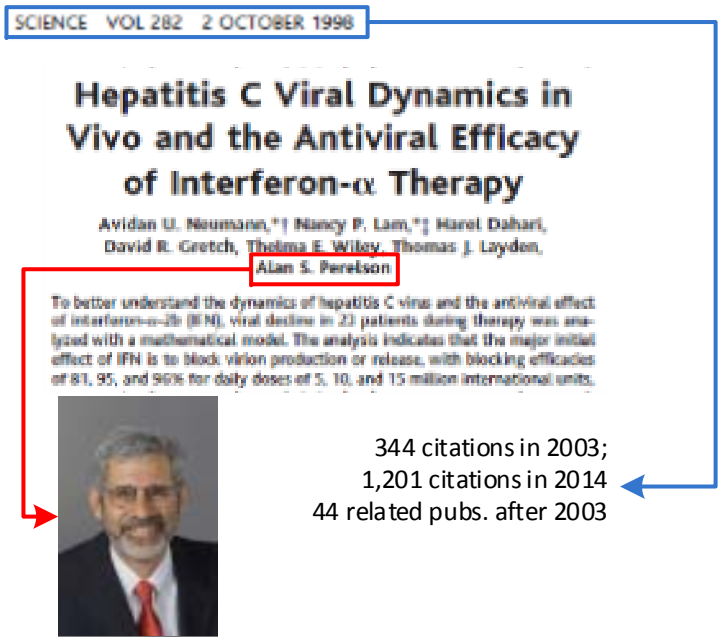
Wooldridge, Jeffrey M. 1997. "Quasi-Likelihood Methods for Count Data." In M. Hashem Pesaran, and Peter Schmidt (Eds.), *Handbook of Applied Econometrics*, pp. 352-406. Oxford: Blackwell.

Figure I: From Source to Related Articles

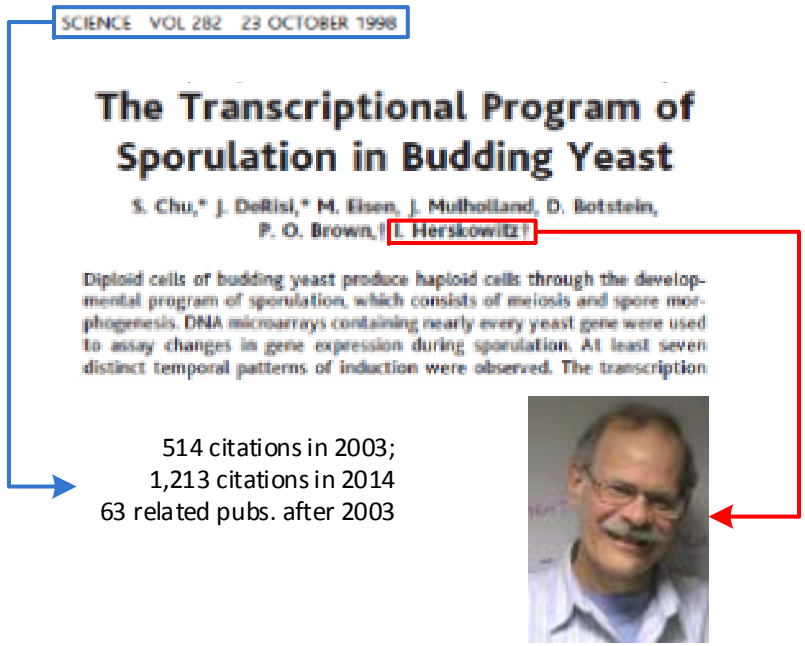


Note: We illustrate the process of identifying the related articles through the use of an example. Ira Herskowitz, a superstar scientist in our sample died in 2003. In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications. One of these publications is *"The transcriptional program of sporulation in budding yeast,"* an article published in the journal *Science* in 1998. On the right-hand side panel, one sees that PMRA identifies 72 related articles related to this source publication. Each of these related articles can then be parsed in a variety of ways. In particular, their authorship list can be matched to the AAMC Faculty Roster, which allows us to distinguish between collaborators of Herskowitz's and non-collaborators, as well as between the subfield's insiders vs. outsiders.

Figure II: Matching Procedure to Identify Controls for the Source Articles



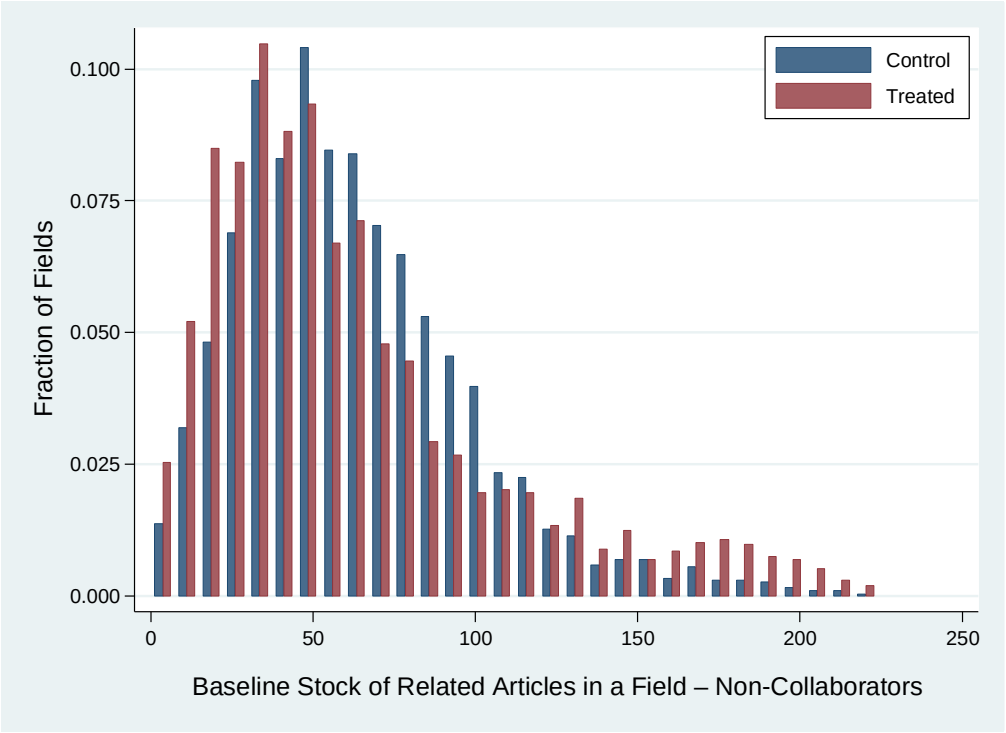
Alan Perelson (1947-)
PhD, 1972
Theoretical Biology & Biophysics, Los Alamos National Lab
In 2003: 173 pubs., 19,817 citations, $7.6 mn. in NIH funding



Ira Herskowitz (1946-2003)
PhD, 1971
Yeast genetics, UCSF
In 2003: 153 pubs., 21,549 citations, $16.8 mn. in NIH funding

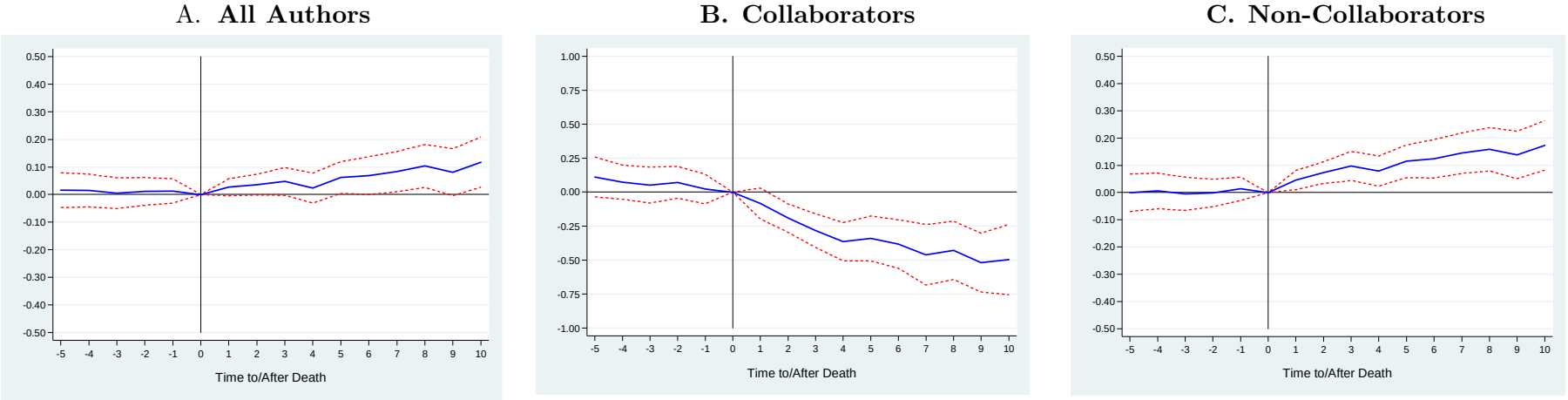
Note: The two articles above illustrate the Coarsened Exact Matching (CEM) procedure (Appendix D provides more details). These two articles appeared in the journal *Science* in 1998. They received a similar number of citations up to the end of the baseline year (2002, one year before Herskowitz's death: 514 citations for Chu et al., 344 citations for Neumann et al. Note that Alan Perelson and Ira Herskowitz are both in last authorship position, They also obtained their PhD within a year of each other.

Figure III: Cumulative Stock of Publication at Baseline



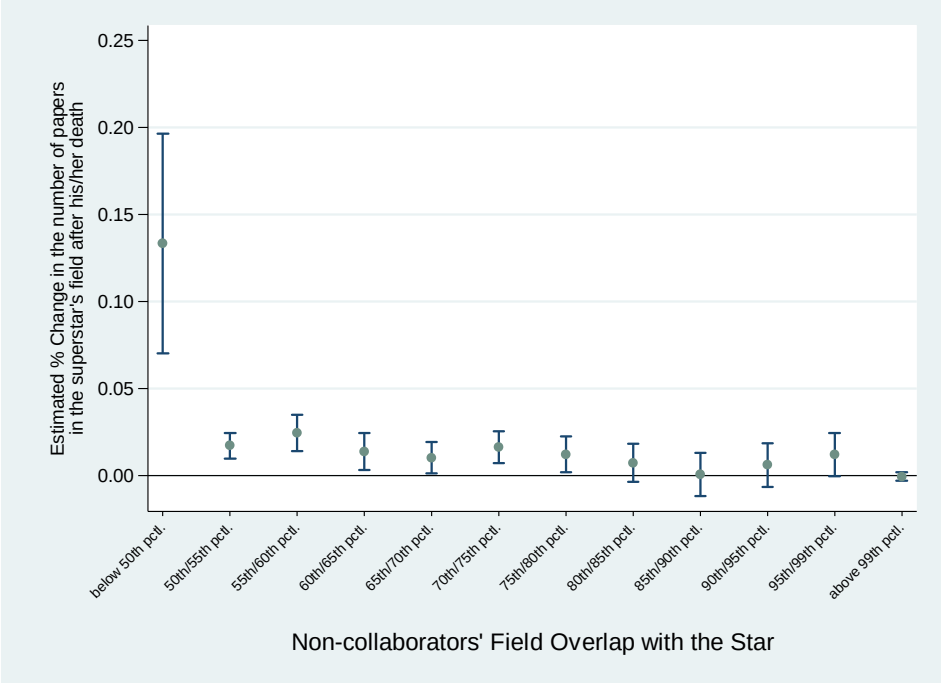
Note: We compute the cumulative number of publications, up to the year that immediately precedes the year of death (or counterfactual year of death), between 3,071 treated subfields and 25,394 control subfields.

Figure IV
Effect of Star Scientist Death of Subfield Growth and Decline



Note: The solid blue lines in the above plots correspond to coefficient estimates stemming from conditional fixed effects quasi-maximum likelihood Poisson specifications in which publication flows in subfields are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/elapsed since the extinction event (the indicator variable for treatment status interacted with the year of death is omitted). The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with dashed red lines; Panel A corresponds to a dynamic version of the specification in column (1) of Table III; Panel B corresponds to a dynamic version of the specification in column (2) of Table III; Panel C corresponds to a dynamic version of the specification in column (3) of Table III.

Figure V: Characteristics of Entering Authors



Note: Each dot corresponds to the magnitude of the treatment effect in a separate regression where the dependent variable is the number of articles in each subfield that belong to a particular intellectual proximity bin. We create a metric of intellectual proximity for each matched author on a related article, by computing the fraction of their publications that belongs to the star's subfield up to its year of publication.

Table I: Summary statistics for extinct superstar scientists (N=451)

	Mean	Median	Std. Dev.	Min.	Max.
Year of Birth	1930.157	1930	11.011	1899	1959
Degree year	1957.633	1957	11.426	1928	1986
Year of Death	1991.128	1992	8.055	1975	2003
Age at Death	60.971	61	9.778	34	91
Female	0.102	0	0.303	0	1
MD Degree	0.403	0	0.491	0	1
PhD Degree	0.489	0	0.500	0	1
MD/PhD Degree	0.108	0	0.311	0	1
Sudden Death	0.409	0	0.492	0	1
Nb. of Subfields	6.801	4	7.298	1	57
Career Nb. of Pubs.	138.221	112	115.704	12	1,380
Career Nb. of Citations	8,341	5,907	8,562	120	72,122
Career NIH Funding	$16,637,919	$10,899,139	$25,441,933	0	$329,968,960
Sits on NIH Study Section	0.007	0	0.081	0	1
Career Nb. of Editorials	0.131	0	0.996	0	17

Note: Sample consists of 451 superstar life scientists who died while still actively engaged in research. See Appendix A for more details on sample construction.

Table II: Summary statistics for control & treated subfields at baseline

	Mean	Median	Std. Dev.	Min.	Max.
Control Subfields(N=25,394)					
Baseline Stock of Related Articles in the Field	75.503	70	40.597	2	232
Baseline Stock of Related Articles in the Field, Non-Collaborators	62.625	57	35.489	1	222
Baseline Stock of Related Articles in the Field, Collaborators	12.877	11	9.710	0	105
Source Article Nb. of Authors	3.969	3	1.792	1	15
Source Article Citations at Baseline	16.307	6	28.023	0	354
Source Article Long-run Citations	70.464	46	93.259	1	1505
Investigator Gender	0.067	0	0.167	0	1
Investigator Year of Degree	1960.546	1962	10.918	1926	1989
Death Year	1991.113	1991	7.965	1975	2003
Age at Death	2.102	2	1.169	0	4
Investigator Cuml. Nb. of Publications	164	142	100	1	861
Investigator Cuml. NIH Funding at Baseline	$18,782,976	$14,268,500	$20,025,386	$0	220,856,880
Investigator Cuml. Nb. of Citations	12,120	9,879	9,960	9	143,383
Treated Subfields (N=3,071)					
Baseline Stock of Related Articles in the Field	75.148	62	51.088	1	237
Baseline Stock of Related Articles in the Field, Non-Collaborators	62.374	50	45.749	0	224
Baseline Stock of Related Articles in the Field, Collaborators	12.774	9	12.612	0	94
Source Article Nb. of Authors	3.986	4	1.907	1	14
Source Article Citations at Baseline	16.668	8	36.309	0	920
Source Article Long-run Citations	70.437	35	180.572	1	6598
Investigator Gender	0.099	0	0.299	0	1
Investigator Year of Degree	1960.141	1961	10.898	1928	1986
Death Year	1991.113	1991	7.965	1975	2003
Age at Death	2.102	2	1.169	0	4
Investigator Cuml. Nb. of Publications	169	143	118	12	1,380
Investigator Cuml. NIH Funding at Baseline	$17,625,556	$12,049,690	$24,878,189	$0	$329,968,960
Investigator Cuml. Nb. of Citations	11,561	8,726	10,186	120	72,122

Note: The sample consists of subfields for 451 extinct superstar life scientists and their matched control subfields. See Appendix D for details on the matching procedure. All time-varying variables are measured in the year of superstar death.

Table III: Main effect of superstar extinction

	Publication Flows			NIH Funding Flows (Nb. of Awards)		
	All Authors	Collaborators Only	Non-Collaborators Only	All Authors	Collaborators Only	Non-Collaborators Only
	(1)	(2)	(3)	(4)	(5)	(6)
After Death	0.022	-0.412**	0.077**	0.018	-0.349**	0.106**
	(0.026)	(0.053)	(0.026)	(0.035)	(0.078)	(0.033)
Nb. of Investigators	6,261	6,260	6,261	6,216	5,779	6,195
Nb. of Fields	34,216	34,211	34,216	33,899	30,317	33,766
Nb. of Field-Year Obs.	1,261,018	1,260,833	1,261,018	1,049,718	938,741	1,045,617
Log Likelihood	-2,785,278	-876,053	-2,631,744	-1,306,848	-516,137	-1,160,093

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year (columns 1, 2, and 3), or the total number of NIH grants that acknowledge a publication in a subfield (columns 4, 5, and 6). All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in column (3) imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.077]\text{-}1)=8.00\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Table IV: Breakdown by long-run citation impact [Non-collaborators only]

	All Pubs	Bttm. Quartile	2nd Quartile	3rd Quartile	Btw. 75th and 95th pctl.	Btw. 95th and 99th pctl.	Above 99th pctl.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: All causes of death							
After Death	0.077**	-0.047	-0.002	0.021	0.113**	0.226**	0.315**
	(0.026)	(0.033)	(0.030)	(0.029)	(0.032)	(0.047)	(0.076)
Nb. of Investigators	6,261	6,203	6,260	6,259	6,257	6,150	5,263
Nb. of Fields	34,216	33,370	34,202	34,211	34,206	33,172	21,579
Nb. of Field-Year Obs.	1,261,018	1,230,048	1,260,506	1,260,833	1,260,648	1,222,550	795,169
Log Likelihood	-2,631,744	-555,616	-1,074,971	-1,399,434	-1,439,621	-525,960	-150,012
Panel B: Anticipated							
After Death	0.108**	0.013	0.052	0.067†	0.133**	0.185**	0.307**
	(0.034)	(0.045)	(0.038)	(0.037)	(0.044)	(0.066)	(0.108)
Nb. of Investigators	4,024	3,970	4,023	4,022	4,020	3,942	3,206
Nb. of Fields	15,104	14,768	15,099	15,102	15,096	14,621	9,464
Nb. of Field-Year Obs.	556,629	544,337	556,444	556,555	556,333	538,812	348,695
Log Likelihood	-1,175,376	-254,163	-483,606	-621,393	-631,853	-227,833	-64,739
Panel C: Sudden							
After Death	0.041	-0.105*	-0.053	-0.033	0.088†	0.266**	0.339**
	(0.042)	(0.051)	(0.049)	(0.048)	(0.052)	(0.070)	(0.109)
Nb. of Investigators	4,654	4,593	4,654	4,654	4,654	4,586	3,758
Nb. of Fields	17,525	17,031	17,516	17,522	17,524	17,035	11,204
Nb. of Field-Year Obs.	645,751	627,657	645,424	645,640	645,714	627,724	412,813
Log Likelihood	-1,322,946	-272,594	-534,918	-706,742	-739,715	-276,358	-79,811

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year, where these publications fall in a particular quantile bin of the long-run, vintage-adjusted citation distribution for the universe of journal articles in *PubMed*. Panel B and Panel C present the same specifications, but run on two distinct subsamples: In Panel B, the 1,576 subfields associated with 229 stars whose death is anticipated (along with the corresponding control subfields); and in Panel C, the 1,342 subfields associated with 185 stars whose death is sudden and unexpected (along with the corresponding control subfields). All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in column (1), Panel A, imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.077]\text{-}1)=8.00\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Table V: Breakdown by intellectual proximity to the work of the star [Non-collab. only]

Panel A	**All Pubs**	**Cardinal Measure**		**Ordinal Measure**	
		Intllct. Proximate Articles	Intllct. Distant Articles	Intllct. Proximate Articles	Intllct. Distant Articles
After Death	0.077**	0.105**	0.061*	0.120**	0.064*
	(0.026)	(0.031)	(0.027)	(0.029)	(0.027)
Nb. of Investigators	6,261	6,102	6,215	6,259	6,261
Nb. of Fields	34,216	30,580	33,786	34,192	34,216
Nb. of Field-Year Obs.	1,261,018	1,126,893	1,245,219	1,260,130	1,261,018
Log Likelihood	-2,631,744	-880,891	-2,287,423	-1,083,451	-2,331,020

Panel B	**In-field vs. out-of-field references**		**Backward citations to the star's bibliome**		**Average backward citation lag**	
	w/ in-field references	w/o in-field references	w/ references to the star	w/o references to the star	Below Median	Above Median
After Death	0.027	0.106**	0.011	0.094**	0.069*	-0.003
	(0.030)	(0.028)	(0.030)	(0.029)	(0.034)	(0.029)
Nb. of Investigators	6,261	6,258	6,261	6,254	6,261	6,260
Nb. of Fields	34,214	34,199	34,214	34,185	34,213	34,213
Nb. of Field-Year Obs.	1,260,944	1,260,396	1,260,944	1,259,883	1,260,917	1,260,923
Log Likelihood	-1,838,530	-1,729,233	-1,917,234	-1,614,955	-1,825,661	-1,708,586

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. In Panel A, the dependent variable is the total number of publications in a subfield in a particular year, where these publications can either be proximate in intellectual space to the star's source publication, or more distant (in the PMRA sense). Since PMRA generates both a cardinal and an ordinal measure of intellectual proximity, we parse the related articles using both measures, yielding a total of four different specifications (the first column of the table merely replicates the estimate already found in Table III, column 3, for comparison purposes. For the cardinal measure, a related article is deemed proximate if its similarity score is above .70, which corresponds to the top quartile of similarity in the sample; For the ordinal measure, a related article is deemed proximate if its similarity rank is below 40, which also corresponds to the top quartile of similarity in the sample. In Panel B, we separate the related articles by examining the type of references cited in their bibliography. Each cited reference can be either in the source's PMRA field, or outside of it; it can be a publication of the star scientist, or of someone else's; and the average lag between the related article's publication year and that of the articles it cites can be either above or below the median (6.5 years). All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in the first column imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.077]\text{-}1)=8.00\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Table VI: Breakdown by related author characteristics [Non-collaborators only]

	Author Career Age		Star Author		Elite Author	
	> 16	≤ 16	with	without	with	without
After Death	0.096**	0.095**	0.022	0.050†	-0.131†	0.068**
	(0.028)	(0.031)	(0.034)	(0.027)	(0.077)	(0.026)
Nb. of Investigators	6,248	6,247	6,247	6,248	5,604	6,248
Nb. of Fields	34,169	34,169	34,146	34,173	27,944	34,173
Nb. of Field-Year Obs.	1,259,281	1,259,281	1,258,436	1,259,429	1,030,092	1,259,429
Log Likelihood	-1,292,332	-1,430,637	-1,295,799	-2,212,397	-308,801	-2,598,287

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year, where these publications have scientists on their authorship roster with certain demographic characteristics. The first two columns examine the impact of related author age. Hence, we compute the average career age of every author we could match with the AAMC Roster, and compute the average age of the authorship team for the related article, at the time of its publication. We then divide related articles according to whether the average career age for identified authors is above or below 16 (the median in our sample), and we aggregate up our measure of subfield activity separately for these two groups. We proceed similarly for the middle two columns (whether or not a related article has one of our 12,935 star on its authorship roster) and for the last two columns (whether or not a related article has a member of the NAS or an HHMI investigator on its authorship roster). All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in the first column imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.096]\text{-}1)=10.08\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Table VII: Breakdown by star scientist characteristics [Non-collaborators only]

	Career Citations		Age at Time of Death		Importance to the Field		Commitment to the Field	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.198†	0.122†	0.103**	0.029	0.042	0.152**	0.060*	0.069†
	(0.109)	(0.074)	(0.039)	(0.033)	(0.027)	(0.041)	(0.030)	(0.037)
Nb. of Investigators	4,486	2,767	5,383	2,705	5,025	4,474	4,231	4,780
Nb. of Fields	16,860	17,356	24,739	9,477	16,978	17,238	15,348	18,868
Nb. of Field-Year Obs.	621,957	639,061	912,578	348,440	625,697	635,321	564,924	696,094
Log Likelihood	-1,334,059	-1,291,548	-1,904,145	-720,086	-1,359,636	-1,233,123	-1,163,783	-1,462,626

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year. Each pair of columns splits the sample across the median of a particular covariate for the sample of fields (treated or control) in the baseline year. For example, the fifth and sixth columns compare the magnitude of the treatment effect for stars that are important to their fields, vs. relatively less important. Importance is defined as the proportion of articles in the subfield that appeared up to the year of death for which the star is an author. In our sample, the median of this covariate is equal to 11.60%. All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in the sixth column imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.152]\text{-}1)=16.41\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Table VIII: Indirect Control and the Regulation of Entry [Non-collaborators only]

Panel A	**Funding**		**Nb. of Trainees**	
	Below Median	Above Median	Below Median	Above Median
After Death	0.182†	0.098	0.101**	0.052
	(0.096)	(0.152)	(0.036)	(0.038)
Nb. of Investigators	4,259	2,957	5,164	2,092
Nb. of Fields	15,487	15,725	18,937	15,279
Nb. of Field-Year Obs.	571,343	579,115	698,154	562,864
Log Likelihood	-1,198,051	-1,200,350	-1,461,474	-1,169,208

Panel B	**Nb. of Frequent Collaborators (5 coauthorships or more)**				**Nb. of "Pivotal" Collaborators (First/Last Authorship Roster Positions)**			
	Editorial Channel		NIH Study Section Channel		Editorial Channel		NIH Study Section Channel	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.262*	0.061	0.188*	0.038	0.290**	0.028	0.265**	-0.012
	(0.105)	(0.054)	(0.074)	(0.088)	(0.099)	(0.053)	(0.093)	(0.037)
Nb. of Investigators	4,408	3,624	5,772	1,958	4,149	4,195	4,434	4,112
Nb. of Fields	18,687	15,529	27,511	6,705	16,806	17,410	17,542	16,674
Nb. of Field-Year Obs.	689,852	571,166	1,014,384	246,634	620,752	640,266	647,354	613,664
Log Likelihood	-1,494,893	-1,128,745	-2,156,618	-472,557	-1,367,384	-1,249,056	-1,414,255	-1,208,943

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year. Each pair of columns splits the sample across the median of a particular covariate for the sample of fields (treated or control) in the baseline year. For example, the first two columns of Panel B compare the magnitude of the treatment effect for stars whose frequent collaborators (five coauthorships or more) have written an above-median number of editorials in the five years preceding the superstar's death, vs. a below-median number of editorials. The next two columns compare the magnitude of the treatment effect for stars with an above-median number of frequent collaborators who sit on NIH study sections (i.e., funding panels) in the five years preceding the superstar's death, vs. a below-median number of such collaborators. The next four columns have an identical structure, except that important collaborators are identified not through the sheer number of coauthorships, but rather by the degree of "intimacy" they imply, using positions at the extremity of the authorship roster as a mark of importance. All models incorporate a full suite of year effects and subfield age effects. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example. the estimates in the first column of Panel B imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.262]\text{-}1)=29.95\%$.

Robust (QML) standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Appendix A:
Criteria for Delineating the Set of 12,935 "Superstars"

Highly Funded Scientists. Our first data source is the Consolidated Grant/Applicant File (CGAF) from the U.S. National Institutes of Health (NIH). This dataset records information about grants awarded to extramural researchers funded by the NIH since 1938. Using the CGAF and focusing only on direct costs associated with research grants, we compute individual cumulative totals for the decades 1977-1986, 1987-1996, and 1997-2006, deflating the earlier years by the Biomedical Research Producer Price Index. We also recompute these totals excluding large center grants that usually fund groups of investigators (M01 and P01 grants). Scientists whose totals lie in the top ventile (i.e., above the 95^{th} percentile) of either distribution constitute our first group of superstars. In this group, the least well-funded investigator garnered $10.5 million in career NIH funding, and the most well-funded $462.6 million.[i]

Highly Cited Scientists. Despite the preeminent role of the NIH in the funding of public biomedical research, the above indicator of "superstardom" biases the sample towards scientists conducting relatively expensive research. We complement this first group with a second composed of highly cited scientists identified by the Institute for Scientific Information. A Highly Cited listing means that an individual was among the 250 most cited researchers for their published articles between 1981 and 1999, within a broad scientific field.[ii]

Top Patenters. We add to these groups academic life scientists who belong in the top percentile of the patent distribution among academics—those who were granted 17 patents or more between 1976 and 2004.

Members of the National Academy of Science and of the Institute of Medicine. We add to these groups academic life scientists who were elected to the National Academy of Science or the Institute of Medicine between 1970 and 2013.

MERIT Awardees of the NIH. Initiated in the mid-1980s, the MERIT Award program extends funding for up to 5 years (but typically 3 years) to a select number of NIH-funded investigators *"who have demonstrated superior competence, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas."* The specific details governing selection vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application must be scored in the top percentile in a given funding cycle.

Former and current Howard Hughes Medical Investigators. Every three years, the Howard Hughes Medical Institute selects a small cohort of mid-career biomedical scientists with the potential to revolutionize their respective subfields. Once selected, HHMIs continue to be based at their institutions, typically leading a research group of 10 to 25 students, postdoctoral associates and technicians. Their appointment is reviewed every five years, based solely on their most important contributions during the cycle.[iii]

[i] We perform a similar exercise for scientists employed by the intramural campus of the NIH. These scientists are not eligible to receive extramural funds, but the NIH keeps records of the number of "internal projects" each intramural scientist leads. We include in the elite sample the top ventile of intramural scientists according to this metric.

[ii] The relevant scientific fields in the life sciences are microbiology, biochemistry, psychiatry/psychology, neuroscience, molecular biology & genetics, immunology, pharmacology, and clinical medicine.

[iii] See Azoulay et al. (2011) for more details and an evaluation of this program.

Early career prize winners. We also included winners of the Pew, Searle, Beckman, Rita Allen, and Packard scholarships for the years 1981 through 2000. Every year, these charitable foundations provide seed funding to between 20 and 40 young academic life scientists. These scholarships are the most prestigious accolades that young researchers can receive in the first two years of their careers as independent investigators.

Appendix B: Linking Scientists with their Journal Articles

The source of our publication data is *PubMed*, a bibliographic database maintained by the U.S. National Library of Medicine that is searchable on the web at no cost.[iv] *PubMed* contains over 14 million citations from 4,800 journals published in the United States and more than 70 other countries from 1950 to the present. The subject scope of this database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. In order to effectively mine this publicly-available data source, we designed PubHarvester, an open-source software tool that automates the process of gathering publication information for individual life scientists (see Azoulay et al. 2006 for a complete description of the software). PubHarvester is fast, simple to use, and reliable. Its output consists of a series of reports that can be easily imported by statistical software packages.

This software tool does not obviate the two challenges faced by empirical researchers when attempting to link accurately individual scientists with their published output. The first relates to what one might term "Type I Error," whereby we mistakenly attribute to a scientist a journal article actually authored by a namesake; The second relates to "Type II error," whereby we conservatively exclude from a scientist's publication roster legitimate articles:

Namesakes and popular names. *PubMed* does not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials, and an optional suffix. This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Inconsistent publication names. The opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffixes (Jr., Sr., 2nd, etc.), or from multiple patronyms due to changes in spousal status.

To deal with these serious measurement problems, we opted for a labor-intensive approach: the design of individual search queries that relies on relevant scientific keywords, the names of frequent collaborators, journal names, as well as institutional affiliations. We are aided in the time-consuming process of query design by the availability of a reliable archival data source, namely, these scientists' CVs and biosketches. PubHarvester provides the option to use such custom queries in lieu of a completely generic query (e.g, `"azoulay p"[au]` or `"graff zivin js"[au]`). As an example, one can examine the publications of Scott A. Waldman, an eminent pharmacologist located in Philadelphia, PA at Thomas Jefferson University. Waldman is a relatively frequent name in the United States (with 208 researchers with an identical patronym in the AAMC faculty roster); the combination `"waldman s"` is common to 3 researchers in the same database.

[iv] `http://www.pubmed.gov/`

A simple search query for `"waldman sa"[au] OR "waldman s"[au]` returns 377 publications at the time of this writing. However, a more refined query, based on Professor Waldman's biosketch returns only 256 publications.[v]

The above example also makes clear how we deal with the issue of inconsistent publication names. PubHarvester gives the end-user the option to choose up to four *PubMed*-formatted names under which publications can be found for a given researcher. For example, Louis J. Tobian, Jr. publishes under `"tobian l"`, `"tobian l jr"`, and `"tobian lj"`, and all three names need to be provided as inputs to generate a complete publication listing. Furthermore, even though Tobian is a relatively rare name, the search query needs to be modified to account for these name variations, as in `("tobian l"[au] OR "tobian lj"[au])`.

Appendix C: *PubMed* Related Citations Algorithm [PMRA]

Traditionally, it has been very difficult to assign to individual scientists, or articles, a fixed address in "idea space," and this data constraint explains in large part why bibliometric analyses typically focus on the determinants of the rate of scientific progress rather than its direction. The empirical exercise in this paper hinges crucially on the ability to relax this constraint in a way that is consistent across extinction events and also requires little, if any, human judgement.

This challenge is met here by the use of the *PubMed* Related Citations Algorithm [PMRA], a probabilistic, topic-based model for content similarity that underlies the "related articles" search feature in *PubMed*. This database feature is designed to aid a typical user search through the literature by presenting a set of records topically related to any article returned by a *PubMed* search query.[vi] To assess the degree of intellectual similarity between any two *PubMed* records, PMRA relies crucially on MeSH keywords. MeSH is the National Library of Medicine's [NLM] controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. There are 27,149 descriptors in the 2013 MeSH edition. Almost every publication in *PubMed* is tagged with a set of MeSH terms (between 1 and 103 in the current edition of *PubMed*, with both the mean and median approximately equal to 11). NLM's professional indexers are trained to select indexing terms from MeSH according to a specific protocol, and consider each article in the context of the entire collection (Bachrach and Charen 1978; Névéol et al. 2010). What is key for our purposes is that the subjectivity inherent in any indexing task is confined to the MeSH term assignment process and does not involve the articles' authors.

Using the MeSH keywords as input, PMRA essentially defines a distance concept in idea space such that the proximity between a source article and any other *PubMed*-indexed publication can be assessed. The algorithm focuses on the smallest neighborhood in this space that includes 100 related records.[vii] The following paragraphs were extracted from a brief description of PMRA:

[v] `((((("waldman sa"[au] NOT (ether OR anesthesia)) OR ("waldman s"[au] AND (murad OR philadelphia[ad] OR west point[ad] OR wong p[au] OR lasseter kc[au] OR colorectal))) AND 1980:2013[dp])`

[vi] Lin and Wilbur (2007) report that one fifth of "non-trivial" browser sessions in *PubMed* involve at least one invocation of PMRA.

[vii] However, the algorithm embodies a transitivity rule as well as a minimum distance cutoff rule, such that the effective number of related articles returned by PMRA varies between 58 and 2,097 in the larger sample of 3,071 source articles published by the 452 star scientists in the five years preceding their death. The mean is 185 related articles, and the median 141.

The neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. To carry out such a program, one must first define what a word is. For us, a word is basically an unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, new lines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next, a limited amount of stemming of words is done, but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles are added in a second time to give them a small advantage in the local weighting scheme. MeSH terms are placed in a third category, and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. If a MeSH term is starred (indicating a major concept in a document), the star is ignored. These three categories of words (or phrases in the case of MeSH) comprise the representation of a document. No other fields, such as Author or Journal, enter into the calculations.

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used, it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) the number of times the term occurs in a particular document; and 3) the number of term occurrences in the document. The first of these pieces of information is used to produce a number called the global weight of the term. The global weight is used in weighting the term throughout the database. The second and third pieces of information pertain only to a particular document and are used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents, its weight is computed as the product of the global weight times the two local weights (one pertaining to each of the documents).

The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about a document. On the other hand, a term that occurred in only 100 documents of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only 10 documents is likely to be even more informative and will receive an even higher weight.

The local weight of a term is the measure of its importance in a particular document. Generally, the more frequent a term is within a document, the more important it is in representing the content of that document. However, this relationship is saturating, i.e., as the frequency continues to go up, the importance of the word increases less rapidly and finally comes to a finite limit. In addition, we do not want a longer document to be considered more important just because it is longer; therefore, a length correction is applied.

The similarity between two documents is computed by adding up the weights of all of the terms the two documents have in common. Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are pre-computed for each document in PubMed so that when one selects Related Articles, the system has only to retrieve this list. This enables a fast response time for such queries.[viii]

Given our set of source articles, we delineate the scientific fields to which they belong by focusing on the set of articles returned by PMRA that satisfy two additional constraints: (i) they are original articles (as opposed to editorials, comments, reviews, etc.); and (ii) they appear in journals indexed by the *Web of Science* (so that follow-on citation information can be collected).

To summarize, PMRA is a modern implementation of *co-word analysis*, a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., title words or phrases, or keywords) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these text (Callon et al. 1989; He 1999). One long-standing concern among practitioners of this technique has been the "indexer effect" (Whittaker 1989). Clustering algorithm such as PMRA assume that the scientific corpus has been correctly indexed. But what if the indexers who chose the keywords brought their own "conceptual baggage" to the indexing task, so that the pictures that emerge from this process are more akin to their conceptualization than to those of the scientists whose work it was intended to study?

[viii] Available at `http://ii.nlm.nih.gov/MTI/related.shtml`

Indexer effects could manifest themselves in three distinct ways. First, indexers may have available a lexicon of permitted keywords which is itself out of date. Second, there is an inevitable delay between the publication of an article and the appearance of an entry in *PubMed*. Third, indexers, in their efforts to be helpful to users of the database, may use combinations of keywords which reflect the conventional views of the field. The first two concerns are legitimate, but probably have only a limited impact on the accuracy of the relationships between articles which PMRA deems related. This is because the NLM continually revises and updates the MeSH vocabulary, precisely in an attempt to neutralize keyword vintage effects. Moreover, the time elapsed between an article's publication and the indexing task has shrunk dramatically, though time lag issues might have been a first-order challenge when MeSH was crated, back in 1963. The last concern strikes us as being potentially more serious; a few studies have asked authors to validate ex post the quality of the keywords selected by independent indexers, with generally encouraging results (Law and Whittaker 1992). Inter-indexer reliability is also very high (Wilbur 1998).

In Table C1, we illustrate the use of PMRA with an example taken from our sample. Ira Herskowitz is a faculty member in our sample who died in 2003. *"The transcriptional program of sporulation in budding yeast"* (*PubMed* ID #9784122) is a publication from his lab which appeared in the October 23^{rd} 1998 issue of the journal *Science* and lists 15 MeSH terms and 5 substances. *PubMed* ID #12242283 is its most related paper according to the PMRA algorithm; it appeared in *Molecular and Cell Biology* in October of 2002 and has 24 MeSH terms (resp. 11 substances). The keywords that overlap exactly have been highlighted in dark blue; those whose close ancestors in the MeSH keyword hierarchical tree overlap have been highlighted in light blue. These terms include common terms such as `Saccharomyces cerevisiae` and `Transcription Factors` as well as more specific keywords including `NDT80 protein, S cerevisiae` and `Gene Expression Regulation, Fungal.`

Table C1: PMRA and MeSH Term Overlap—An Example

Source Article	PMRA-Linked Article
Chu et al., "The transcriptional program of sporulation in budding yeast." *Science*, 1998.	Sopko et al. "Phosphorylation and maximal activity of Saccharomyces cerevisiae meiosis-specific transcription factor Ndt80 is dependent on Ime2." *MCB*, 2002.
PMID #9784122	**PMID #12242283**
MeSH Terms	**MeSH Terms**
Animals	Active Transport, Cell Nucleus
Chromosomes, Fungal	Binding Sites
DNA-Binding Proteins*	Cell Cycle Proteins*
Fungal Proteins	Cell Nucleus
Gene Expression Regulation, Fungal*	DNA-Binding Proteins*
Genes, Fungal	Fungal Proteins*
Genome, Fungal	Gene Expression Regulation, Fungal*
Humans	Genes, Fungal
Meiosis	Intracellular Signaling Peptides and Proteins
Morphogenesis	Meiosis*
Organelles	Phosphorylation
Saccharomyces cerevisiae*	Promoter Regions, Genetic
Spores, Fungal	Protein Kinases*
Transcription Factors	Protein-Serine-Threonine Kinases
Transcription, Genetic*	Recombinant Fusion Proteins
	Saccharomyces cerevisiae
	Saccharomyces cerevisiae Proteins*
	Spores, Fungal
	Substrate Specificity
	Transcription Factors*
	Transcriptional Activation
Substances	**Substances**
DNA-Binding Proteins	Cell Cycle Proteins
Fungal Proteins	DNA-Binding Proteins
NDT80 protein, S cerevisiae	Fungal Proteins
Saccharomyces cerevisiae Proteins	Intracellular Signaling Peptides and Proteins
Transcription Factors	NDT80 protein, S cerevisiae
	Recombinant Fusion Proteins
	Saccharomyces cerevisiae Proteins
	Transcription Factors
	Protein Kinases
	IME2 protein, S cerevisiae
	Protein-Serine-Threonine Kinases

Appendix D: Construction of the Control Group

We detail the procedure implemented to identify the control subfields that help pin down the life-cycle and secular time effects in our difference-in difference (DD) specification. Happenstance might yield a sample of stars clustered in decaying scientific fields. More plausibly, activity in the typical subfield might be subject to idiosyncratic life-cycle patterns, with their productive potential first increasing over time, eventually peaking, and thereafter slowly declining. Relying solely on subfields treated earlier or later as as an implicit control group raises the worry that these time-varying omitted variables will not be fully captured by subfield age controls, particularly since dating the birth of a subfield is a process fraught with hazards.

To address this concern, we create an additional level of difference by selecting control subfields. Recall that selecting a subfield in our framework is akin to first selecting a source article and then using PMRA to harvest all the related articles to this source in intellectual space. Since the second step is fully automated, only the first step is really of concern. Practically, we will recruit control source articles from the set of articles authored by star scientists who do not die prematurely. But what makes a satisfactory control group? It is important to distinguish between *ex ante* vs. *ex post* criteria. *Ex ante*, one would like control source articles to have the following properties:

1. to be published contemporaneously with the source article for the treated subfield;
2. to be unrelated in both an intellectual and a social sense, to the source article for the treated subfield;
3. to be of similar expected impact and fruitfulness, relative to the source article for the treated subfield;
4. to have a similar number of authors as the source article for the treated subfield;
5. to have a superstar author in the same authorship position and of approximately the same age as that occupied by the extinct superstar on the authorship roster of the source article for the treated subfield.

Ex post, it will be important for the control subfields to satisfy an additional condition: the treated and control subfields should exhibit very similar trends in publication activity and funding flows up to the year of treatment (i.e., the year of death for the treated superstar).

Coarsened Exact Matching. To meet these goals, we implement a "Coarsened Exact Matching" (CEM) procedure (Blackwell et al. 2009). The first step is to select a relatively small set of covariates on which we need to guarantee balance *ex ante*. This choice entails judgement, but is strongly guided by the set of criteria listed above. The second step is to create a large number of strata to cover the entire support of the joint distribution of the covariates selected in the previous step. In a third step, each observation is allocated to a unique strata, and for each observation in the treated group, control observations are selected from the same strata.

The procedure is coarse because we do not attempt to precisely match on covariate values; rather, we coarsen the support of the joint distribution of the covariates into a finite number of strata, and we match a treated observation if and only if a control observation can be recruited from this strata. An important advantage of CEM is that the analyst can guarantee the degree of covariate balance *ex ante*, but this comes at a cost: the more fine-grained the partition of the support for the joint distribution (i.e., the higher the number of strata), the larger the number of unmatched treated observations.

Implementation. We identify controls based on the following set of covariates (t denotes the year of death): star scientist career age, citations received by the article up to year t, number of authors; position of the star

author on the authorship roster (only first or last authorship positions are considered); journal; and year of publication. The first three covariates only need to match within relatively coarse bins. For instance, we create nine career age categories: less than 10 years; between 10 and 20 years; between 20 and 25 years; between 25 and 30 years; between 30 and 35 years; between 35 and 40 years; between 40 and 45 years; between 45 and 50 years, over 50 years of career age. Similarly, we coarsen the distribution of citations at baseline into five mutually exclusive bins: zero citations; between one and 10 citations; between 10 and 50 citations; between 50 and 120 citations; and more than 120 citations. In contrast, we impose an exact match on journal, publication year, and the star's authorship position.

We match approximately 75% of the treated source articles in this way. Unfortunately, some further trimming of the control articles is needed. First, we eliminate any control that shares any author with the treated source. Second, we eliminate any control article with a dead star scientist on its authorship roster, even if s/he appears in an intermediate position in the authorship list. Third, we drop every control that also happens to be related intellectually to its source as per PMRA. Finally, we drop from the data any source article that finds itself an orphan (i.e., not paired with any control) at the conclusion of this process. Figure II provides an illustrative example.

The final sample has 3,071 treated source articles and 31,102 control source articles. As can be seen in Figure III, the distribution of activity levels, measured by cumulative publications up to the baseline year, is very similar between treated and control subfields. As well, there is no evidence of preexisting trends in activity, as demonstrated by the coefficient estimates graphed in Figure IV. In Table II, treated and control subfields are very well-balanced on the covariates that formed the basis of the CEM matching procedure. This is true almost by construction. What is more surprising (and also welcome) is that the procedure balances a number of covariates that were not used as inputs for matching, such as various metrics of star eminence. For other covariates, we can detect statistically significant mean differences, though they do not appear to be substantively meaningful (e.g., 6.7% of control stars vs. 9.9% of treated stars are female).

Sensitivity Analyses. Human judgement matters for the outcome of the CEM procedure insofar as she must draw a list of "reasonable" covariates to match on, as well as decide on the degree of coarsening to impose. We have verified that slight variations in the implementation (e.g., varying slightly the number of cutoff points for the stock of baseline citations for the source; focusing on birth age as opposed to career age for the stars) have little impact on the main results.

References

Azoulay, Pierre, Andrew Stellman, and Joshua Graff Zivin. 2006. "PublicationHarvester: An Open-source Software Tool for Science Policy Research." *Research Policy* **35**(7): 970-974.

Azoulay, Pierre, Joshua Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* **42**(3): 527-554.

Bachrach, C. A., and Thelma Charen. 1978. "Selection of MEDLINE Contents, the Development of its Thesaurus, and the Indexing Process." *Medical Informatics (London)* **3**(3): 237-254.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "`cem`: Coarsened Exact Matching in Stata." *The Stata Journal* **9**(4): 524-546.

Callon, Michel, Jean-Philippe Courtial, and F. Laville. 1991. "Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry." *Scientometrics* **22**(1): 155-205.

He, Qin. 1999. "Knowledge Discovery Through Co-Word Analysis." *Library Trends* **48**(1): 133-159.

Law, John, and John Whittaker. 1992. "Mapping Acidification Research: A Test of the Co-word Method." *Scientometrics* **23**(3): 417-461.

Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* **8**(423).

Névéol, Aurélie, Rezarta Islamaj Dogan, and Zhiyong Lu. 2010. "Author Keywords in Biomedical Journal Articles." *AMIA Symposium Proceedings* 537-541.

Whittaker, John. 1989. "Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis." *Social Studies of Science* **19**(3): 473-496.

Wilbur, W. John. 1998. "The Knowledge in Multiple Human Relevance Judgments." *ACM Transactions on Information Systems* **16**(2): 101-126.