Research Policy 44 (2015) 1629–1632

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

The New Data Frontier
Special issue of Research Policy

Maryann Feldman[a,*], Martin Kenney[b], Francesco Lissoni[c]

[a] University of North Carolina, Chapel Hill, USA
[b] University of California, Davis, USA
[c] GREThA-Université de Bordeaux CRIOS-Università Bocconi, Italy

ARTICLE INFO

Article history:
Received 23 February 2015
Accepted 23 February 2015
Available online 15 April 2015

Keywords:
Digital data
Administrative records
Social science research

We are now in the age of Big, and, seemingly, ever Bigger Data. The current public discussion focuses on the avalanche of data, due to fact that nearly all written (and other) materials are now available in a digital format, which simplifies their accessibility, extraction, classification, and analysis. Even more so, the adoptions of online digital platforms create new and ever-larger data quantities every day. While created for other purposes the potential for scientific socio-economic research appears simultaneously extremely promising and extremely uncertain – very much like answers in search of good questions. Amidst the great hype, there are continuing controversies about how to define and delineate Big Data. Because of these ambiguities and questions, we choose not to use the phrase "Big Data" for this Special Issue, but instead focus on the new data frontiers that scholars in our research community might find useful. While the picture is still evolving in terms of what information will be useful and which tools are most efficient for accessing and manipulating data, what is certain is that changes during the last decade enabled by new technologies have dramatically enhanced the availability, scale and ability to connect previously disparate data sources. These existing and emerging data sources provide new opportunities to address questions of interest to Research Policy readers and to demonstrate the potential for providing previously elusive empirical evidence.

Changes in access to information and data during the past thirty years have been remarkable. To illustrate, as undergraduates, Feldman and Kenney relied on punch cards, computer time was valuable and running a simple regression required writing code and waiting hours only to find a punctuation error in programming syntax. Data was available on magnetic tape and requiring arcane Job control language to access remote mainframe computers. For his Master's dissertation, Lissoni crunched data at an IBM/VM terminal and word processed on a Mac with just a 3.5-inch floppy drive and no hard disk, creating tables by meticulously copying numbers from printouts. Now remarkable volumes of data are easily accessible electronically, not only increasing the number of observations available, but, more dramatically, changing the types of questions that can be addressed.

E-government administrative records, proprietary data sources and social media are now potential sources of data for analysis. There are remarkable opportunities available as data from initially incompatible sources can be matched and combined in relational databases to examine associations that were previously evasive. The net result is great opportunity to increase our understanding of the patterns of interaction in scientific and inventor networks, and the factors and relationships that define the geography of innovation. The potential is to understand the processes and factors, and to increase the economic value of scientific activity. We are certain that wide dissemination of the potential of new data will contribute to the scholarly understanding of vital unresolved issues in the study of innovation and the process of invention. This is a time when information technology terms such as data disambiguation, entity resolution, or data linkage are entering many social scientists' current vocabulary, pointing to the importance that a number of technical issues now have when it comes not just to producing a

* Corresponding author. Tel.: +1 919 357 4043.
E-mail address: maryann.feldman@gmail.com (M. Feldman).

http://dx.doi.org/10.1016/j.respol.2015.02.007

new dataset, but also to interpreting the results obtained through its use (Varian, 2014).

The Innovation Studies community of scholars, for which Research Policy is the reference journal (Fagerberg and Verspagen, 2009; Fagerberg et al., 2012), has a long tradition in collecting ambitious and creative data as an integral part of the researcher's job. A classic example in this respect was the SPRU innovation database (Robson et al., 1988; Pavitt et al., 1987, 1989), but also Griliches' (1984) pioneering work on patents or the Yale innovation survey (Levin et al., 1985). Recent examples published in Research Policy include Azoulay et al. (2007), Li et al. (2014), Franzoni and Sauermann (2014), among others. And in between, we have observed one of the greatest advances in the field, namely the digitization of patent data, first with the NBER database (Hall et al., 2001) then in Europe, where the collaboration between innovation scholars and the EPO (European Patent Office) has led to the creation of the PatStat platform and community (Giuri et al., 2007). The wide dissemination of this data increased the number of researchers that could contribute to the scholarly understanding of invention.

Pessimists allege that Big Data may bring an end to social science research. One fear is that scholars will focus on pattern recognition rather than developing theory or engaging in hypothesis driven empirical research. As it becomes easier to manipulate large numbers of records it is seductive to keep collecting more and more observations, matching ever more and more diverse sources – the potential is unlimited. Resources may be diverted to never-ending data projects rather than focusing on questions that are answerable with currently available data. Moreover, with a sufficiently large sample it is simply easier to find associations and make dubious claims.[1] Another worry is that rather than focusing on interesting questions researchers will limit their inquiry to questions they are able to examine rather than consider the more socially relevant questions, becoming like the proverbial drunk who seeks their car keys under the lamp post because it is easiest to look there. In the rush to collect larger datasets (and as Conti and Liu (in this issue) note, no scholar would argue against more data points), there is a need to remember context and structure. Rather than become slaves to the concept of Big Data there is an opportunity to expand our inquiry and understanding – to explore a new frontier.

To encourage the use of new data this Special Issue includes articles on new data mining software tools, database disambiguation, and network-mapping analysis. We also include papers that draw upon previously unavailable datasets obtained from digital and internet sources that include sources as diverse as governmental filings, administrative records, proprietary data sources, and social networking sites. Each article details the data collection methodology in greater detail than normally might be provided in a research article to encourage transparency and replicability. Authors were asked to explicitly consider data quality control procedures and data accessibility so that others could benefit from and extend these efforts. The use of new data sources in this Special Issue provide greater granularity and reveal new findings that challenge conventional wisdom. The topics covered by the papers have all attracted a good deal of attention, both by Research Policy and its readers' community at large. Content-wise, they range from the economics and economic sociology of science to industrial dynamics, geography, and entrepreneurship.

The type of data extracted and used include both some classic sources, such as scientific publications and patents as well as much more experimental sources, such as curricula vitae (CVs) or website and social media content. Other sources include administrative records of companies, universities and research funding agencies. Very often, the focus of the analysis is on individuals, whether scientists, inventors or entrepreneurs, whose activity is related to their professional condition, age, gender, ethnicity, and location. In this respect, a key technical issue is name disambiguation, which is at the same time necessary and one that hides many traps (as discussed at length in Ventura et al. (this issue)). Another important technique used in the Special Issue is text mining. So, when it came to choose the ordering of papers, we decided on a mix of all criteria: contents, data, and methods. As for impact and relevance, we leave to readers to decide.

The lead paper by Annamaria Conti and Christopher Liu describes a novel database created to study laboratories as production units, specifically focusing on lab personnel composition. Based on annual reports from MIT's Department of Biology for the years 1966–2000 the authors construct a relational database with detailed data on all laboratory members, including rank, function and role (i.e., graduate student, postdoctoral associate, technician, etc.). Personnel rosters were supplemented with data on publications from the Medline database. From the point of view of the Special Issue, this research is an exemplar of discovering a unique source for building longitudinal datasets and matching them with the more traditional datasets. Conti and Liu's approach addresses micro-level research laboratory organizational issues, including documenting the increased employment of post-docs over 30 years. They find that post-docs with external funding (fellowships) make greater contributions to the laboratory's publication outcomes, while having technicians in the lab is important to producing high-impact publications. Further and quite importantly, the results suggest that constructing laboratory composition from publication author lists may lead to biased findings.

The paper by Aldo Geuna, Rodrigo Katalishi, Manuel Toselli, Eduardo Guzman, Cornelia Lawson, Ana Fernandez-Zubieta and Beatriz Barros, extends the opportunity to study scientific careers by describing a methodology and software tool useful to building a database on the careers and productivity of academics based on CVs. The methodology, Science in Society Observatory (SiSOB), uses data crawling and text mining. Their test case is a sample of 360 US scientists funded by the National Institute of Health (NIH) and 291 UK scientists funded by the Biotechnology and Biological Sciences Research Council (BBSRC), however the tool can be applied to any context where CVs are available in English. The software is available under the free software GNU General Public License. A most striking application of CV-based information concerns the study of scientists' mobility (in space and across organizations). Most existing studies exploit biblio-metric and techno-metric data, but the use of datasets consisting of publications and (especially) patents are based upon discrete events in between which mobility can occur and not be recorded (so that data points for the individuals are too few and thus impacted by various types of censoring and truncations). CV-based observations fill the gap, and allow for far more nuanced views of changes in tasks and roles that may occur with mobility.[2]

Julia Lane, Jason Owen-Smith, Rebecca Rosen, and Bruce Weinberg describe the U-Metrics database, which is based on university transaction-level information about wage and vendor payments from federal research grants. This longitudinal dataset, building upon the US government investment in STAR Metrics provides data for evaluating the process, products, and impact of research. Once again, individuals take center stage, as the focal points to which all the available information must be linked. At the same time, teams and laboratories can be analyzed as network sets. The authors highlight a number of challenges associated with this type of exercise,

[1] For example, see the saga of the Google Flu prediction algorithm (Lazer et al., 2014).

[2] While this research project did not used data sources such as LinkedIn, the editors believe this provides intriguing new possibilities for career studies.

many of which are not purely technical, but also conceptual. In particular, taxonomies and classifications have to be produced, which not only have to follow a logic of convenience (for data retrieval and linkage), but also be informative of the functions and responsibilities of the various individuals in a team or lab, and of the same individuals during their careers.

Several papers address name disambiguation as a key technical issue when creating or matching large datasets. Essentially, this consists of assigning unique identifiers to individuals (or other observation units) that take into consideration all of the name variants and special cases across different datasets. Existing techniques struggle with finding parsimonious ways to deal with problems of precision (minimization of false positives) and recall (minimization of false negatives), both of which may have dramatic consequences for the measurement of individual productivity and positioning in networks of authors and inventors as well as of overall properties of networks (see, for example, Raffo and Lhuillery, 2009; Li et al., 2014). The article by Samuel Ventura, Rebecca Nugent, and Erica Fuchs introduces a learning algorithm trained on hand-disambiguated labeled data. The technique is described and then compared to other patent disambiguation methods. The results suggest that their supervised learning approach dramatically reduces error rates. The code, which allows users to implement the supervised learning approach as well as the test dataset on optoelectronic inventors, is publicly available.

Returning to content, an emerging literature deals with patent-publication pairs – instances of simultaneous disclosures of research results in both a scientific publication and a patent (Gans et al., 2013; Lissoni et al., 2013). Research based on patent-publication pairs allows the investigation of true extent of anti-commons effects due to intellectual property protection of research tools, provided that one finds the way to identify such pairs through large-scale, automated methods (and not through small-scale, expert-based surveys, as in the pioneering paper by Murray and Stern (2007)). Magerman, van Looy, and Debackere explore whether involvement in patenting hampers the dissemination of a scientist's published research. To address this question, they apply their own text-mining algorithms to a dataset consisting of 948,432 scientific publications and 88,248 patents. They identify 584 patent-paper publication pairs. They define a comparison control group of publications without an equivalent patent and then compare the forward citations patterns, testing whether publications in the pairs receive fewer citations than those in a control sample. They found that the publications linked to a patent actually receive more citations than publications without a link to a patent thereby suggesting patenting does not hamper the dissemination of published research. Of particular interest in this paper is the use of text-mining software to identify the patent-paper pairs. The results suggest that there are not any significant anti-commons effects.

Max Nathan and Anna Rosso develop a novel sector-product approach to map industrial activity not captured by the standard industrial classification system, specifically using the example of the size of the UK information economy. To develop a more accurate description of new and emerging industries they begin with a company-level database drawn from UK administrative data. They derive information from unstructured sources produced by proprietary sources and then use sophisticated computational data-cleaning strategies that include text mining to develop estimates that suggest that the UK digital economy is 40% larger and may employ twice as many people as estimated using the standard industrial classification systems. Their results invite speculation along two questions: First, whether in the new digital economy, which is increasingly characterized by ephemeral "firms" and contingent work, measurement of traditional categories such as firms and employment is becoming more difficult. Second, whether new datasets drawing upon various new data sources can provide better insight into important economic policy questions than can single-sourced data or government-collected statistics.

Namil Kim, Hyeokseong Lee, Wonjoon Kim, Hyunjong Lee, and Jonghwan Seo also explore the definition of industrial activity, specifically the blurring of boundaries between previously distinct industries occurring with the functional integration of products and services. To examine this question, they conduct a means co-occurrence-based analysis by text mining over 4 million newspaper articles published from 1989 to 2012 that mentioned firms listed on the NYSE, NASDAQ, or AMEX stock markets. The authors' text-mining technique targeted firm name co-occurrence in a single sentence, and used various techniques also based on specific algorithms for excluding random co-occurrences. The findings suggest that while industry convergence is generally under way, the trends and patterns are quite heterogeneous and sector specific. For Research Policy readers, this article suggests that there are new computational techniques that permit the exploitation of much larger and previously unused data sources and illustrates the potential of text-mining as a tool for tapping in non-academic, non-patent literature as a source of information on industrial dynamics.

Roberto Catini, Dmytro Karamshuk, Orion Penner, and Massimo Riccaboni introduce a new data-driven methodology to define geographic clusters, one that does not rely on exogenous border setting, with a specific application to biotechnology using the geographic distributions of scientific publication output from the PubMed database. Using nine million articles, they match the authors' institutional affiliation data to obtain addresses, which they then geocode to location. With this they are able to identify specific biomedical research clusters within cities. This provides a new methodology to endogenously identify clusters based on research institution location in scientific and technological production at different geographic scale and maps cluster structure emergence and evolution.

The final four articles describe datasets based upon new data sources, and techniques used for building them. This work has at its origin the desire to investigate core issues in science and innovation. In these efforts longitudinal data are organized in relational databases that allow for data linkage among different data elements and provide for the development of rigorous conceptual and empirical models.

Floortje Alkemade, Gaston Heimeriks, Antoine Schoen, Lionel Villard, and Patricia Laurens introduce a unique database Corporate Invention Board (CIB), which combines data from the PATSTAT database with financial data from the ORBIS database for nearly 2300 firms that make the largest R&D investments for the period 1993–2005. The research finds significant national and sectoral heterogeneity of R&D internationalization. The authors find that while national-level indicators explain a large part of the variance observed in the ability of countries to attract R&D from foreign multinationals, there are significant differences between sectors, which has significant implications for the design of foreign R&D and innovation policies.

Martin Kenney and Donald Patton describe an open access dataset of approximately 2000 U.S. emerging growth firms (EGFs) that made an initial stock offering on US public markets from 1990 to 2010. The dataset includes firm descriptors, the location of the firm and its key backers, and various characteristics of the management team and the board of directors. To illustrate possible uses of the dataset, they analyze the gender and nationality of the firms' top management teams and board of directors (approximately 40,000 individuals). Using undergraduate education, as a proxy for nationality, they find that there are more European than Asian immigrants in top management ranks, a finding that, while not squaring with the evidence on migration of scientists and engineers (Kerr, 2008; Freeman, 2014), is in line with more general data on highly skilled migration (Docquier and Rapoport, 2012). These results suggest that the database can be used by various social scientists in stand-alone analyzes, but even more important it can be merged with yet other datasets to address other social scientific questions.

Maryann Feldman and Nichola Lowe describe a longitudinal relational database dedicated to studying the emergence and development of regional economies. The database is organized around firm entry into the region, with a specific focus on technology intensive firms. Details on company founders, annual firm employment and sales, patenting and trademark activity along with engagement in the entrepreneurial ecosystem are tracked using data from a variety of sources disambiguated on firm names and aliases. Their study traces the industrial genesis of technology-intensive entrepreneurial firms in North Carolina's Research Triangle Park and the adjacent area. The paper's primary objective is to describe a transferable framework for analyzing regional dynamics in other locations. In addition to the quantitative data the database is supplemented with archival materials and oral histories with firm founders, corporate executives and institutional actors to provide historical context.

Concluding remarks

This special issue of Research Policy provides readers with an extensive overview of where the New Data Frontier lies. It is not a straight line, one that advances uniformly in the same direction. It dashes forwards into the (so far) alien territories of large-scale data linkage and text mining, but also suggests that digitization of previous records kept only in hard copy are providing previously hard to access records, thereby enriching small-scale, in-depth data collection.[3] Individuals emerge as important observational units. Though this does not imply resurrecting outdated views of science and technology as personal enterprises, it does give us increasing granularity in our research on organizations and networks. In fact, this granularity provides the means to reconsider organizations (whether laboratories or firms) as relational entities, whose boundaries (and their evolution in time) have to be determined contextually in accordance with our theoretical perspective and the hypotheses being tested. It also raises interesting questions concerning the organization of socio-economic research, more generally, and for those of us in the Innovation Studies community, more specifically. On one side, the size of the initial investment in data collection and the sheer size of the tasks to be performed force upon us an increasing division of labor. Our research is becoming more and more team-based, not only in terms of the average number of authors per paper, but also in terms of the division of labor and who should receive credit for the intellectual activity. On the other side, it forces us to reconsider the importance of data commons and sharing, the sole means to avoid duplications of extremely expensive data collection efforts, and to allow for extensive, peer-reviewed data quality checking, through their use and re-use. It will be up to learned societies and leading journals, such as Research Policy, to provide the necessary coordination.

References

Azoulay, P., Ding, W., Stuart, T., 2007. The determinants of faculty patenting behavior: demographics or opportunities? J. Econ. Behav. Org. 63 (4), 599–623.

Docquier, F., Rapoport, H., 2012. Globalization, brain drain, and development. J. Econ. Lit., 681–730.

Fagerberg, J., Verspagen, B., 2009. Innovation studies—the emerging structure of a new scientific field. Res. Policy 38 (2), 218–233.

Fagerberg, J., Fosaas, M., Sapprasert, K., 2012. Innovation: exploring the knowledge base. Res. Policy 41 (7), 1132–1153.

Franzoni, C., Sauermann, H., 2014. Crowd science: the organization of scientific research in open collaborative projects. Res. Policy 43 (1), 1–20.

Freeman, R.B., 2014. Immigration, International Collaboration, and Innovation: Science and Technology Policy in the Global Economy (No. w20521). National Bureau of Economic Research.

Gans, J.S., Murray, F.E., Stern, S., 2013. Contracting over the Disclosure of Scientific Knowledge: Intellectual Property and Academic Publication (No. w19560). National Bureau of Economic Research.

Griliches, Z. (Ed.), 1984. R&D, Patents and Productivity. National Bureau of Economic Research Project Report, University of Chicago Press.

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gam- bardella, A., Garcia-Fontes, W., Geuna, A., Gonzales, R., Harhoff, D., Hoisl, K., Lebas, C., Luzzi, A., Magazzini, L., Nesta, L., Noma- ler, O., Palomeras, N., Patel, P., Romanelli, M., Verspagen, B., 2007. Inventors and invention processes in Europe. Results from the PatVal-EU survey. Res. Policy 36, 1107–1127.

Hall, B.W., Jaffe, A.B., Trajtenberg, M., 2001. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools, NBER Working Paper No. 8498.

Kerr, W.R., 2008. Ethnic scientific communities and international technology diffusion. Rev Econ. Stat. 90 (3), 518–537.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The parable of google flu: traps in big data analysis. Science 343 (14), 1203–1205.

Levin, R.C., Cohen, W.M., Mowery, D.C., 1985. R&D appropriability, opportunity, and market structure: new evidence on some Schumpterian Hypotheses. Am. Econ. Rev. 75 (2), 20–24.

Li, G.C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Fleming, L., 2014. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Res. Policy 43 (6), 941–955.

Lissoni, F., Montobbio, F., Zirulia, L., 2013. Inventorship and authorship as attribution rights: an enquiry into the economics of scientific credit. J. Econ. Behav. Org. 95, 49–69.

Murray, F., Stern, S., 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. J. Econ. Behav. Org. 63 (4), 648–687.

Pavitt, K., Robson, M., Townsend, J., 1987. The size distribution of innovating firms in the UK. 1945–1983. J. Ind. Econ., 297–316.

Pavitt, K., Robson, M., Townsend, J., 1989. Technological accumulation, diversification and organisation in UK companies, 1945–1983. Manage. Sci. 35 (1), 81–99.

Raffo, J., Lhuillery, S., 2009. How to play the names game: patent retrieval comparing different heuristics. Res. Policy 38 (10), 1617–1627.

Robson, M., Townsend, J., Pavitt, K., 1988. Sectoral patterns of production and use of innovations in the UK. 1945–1983. Res. Policy 17 (1), 1–14.

Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28 (2), 3–28.

[3] The editors wish to point out that The Wayback Machine is a treasure trove of historical Internet website records going back to 1996.