Enabling Next Generational Social Science with Machine Reading

Scott Appling
Georgia Institute of Technology
Atlanta, GA
scott.appling@gtri.gatech.edu

Erica Briscoe
Georgia Institute of Technology
Atlanta, GA
erica.briscoe@gtri.gatech.edu

ABSTRACT

The social science research process has traditionally required researchers to engage in a largely manual information seeking process and then manual analysis to extrapolate trends from past work into the study design process including hypotheses generation and variable declaration. Across several computational disciplines including probabilistic relational learning and machine reading, we see opportunity to advance and significantly positively change the social science research process in a world with more and more scientific textual data accruing on a yearly, if not, daily basis. Here we present an articulation of the problem we see with the nature of publishing scientific findings in largely unstructured natural language text along with our perspective for how both micro- and macro-reading methods can play a role together with the work being done on the scientific research cycle itself to drive better and more efficient research across all of science.

CCS CONCEPTS

• **Information systems** → *Information systems applications*; *Data mining*; • **Computing methodologies** → *Natural language processing*; *Information extraction*;

KEYWORDS

Science of Machine Reading

1 INTRODUCTION

The social science research process, and more generally, the scientific research process is a general set of steps, forming a cycle, that researchers within the social sciences generally take as they engage in and conduct research in their sub-fields of interest. The process usually starts in the **model** step (See Figure 1 for our working definition of this process) with one or more questions of scientific inquiry that a researcher wants to formally investigate where the research begins considering prior literature and scaffolding hypotheses; this is seen as the start of a research cycle. These 'investigations' take many forms (e.g. qualitative, quantitative, theoretical, conceptual) and sub-types (e.g. causal, non-causal). Depending on the type of investigation, for example, an experimental design with hypotheses and analyses testing the effects of an independent variable on a dependent variable, different levels of background context are needed by the researcher to appropriately design such a study.

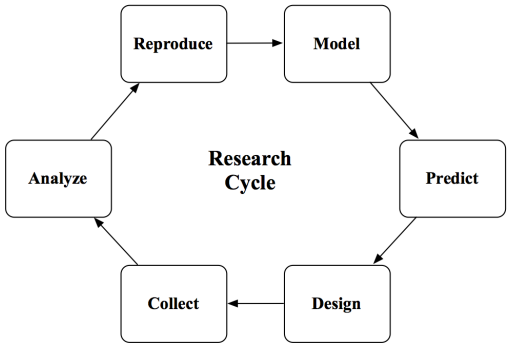

Figure 1: Research Cycle

2 THE PROBLEM

The research process itself, conceived and refined over hundreds of years, typically allows for new research to be designed and conducted by building off of past knowledge. It is however within the past 60 years that the sheer magnitude of the scientific data being observed and collected has resulted in an inability for researchers to keep up and fully utilize it all. Perhaps as a symptom of this or as the global workforce has slowly shifted away from physical labor jobs towards those of science and engineering, the speed of scientific literature growth every year has been rapidly increasing; whereas, the amount of time researchers have to discover, digest, and synthesize new research directions has not been increasing. [5] The state of the research process is such that individual researchers are stuck with the massive data dilemma like professionals in other STEM fields. As this happens, the ability to conduct future research begins to suffer from different kinds of problems e.g. those related to information seeking behaviors [8] or those related to the ways experiment designs are constructed [4].

Researchers are often times left between choosing what appears within the first couple pages of their search platform's results and spending vast amounts of time trying to discover related terms (and consequently, studies) that should likely be considered as a part of their literature review and hypotheses and experiment planning activities. Figure 2 is but one example of a bibliometric database's growth over the past several years; overall there is an increase from year to year as more research publications are produced. Albeit, in recent years there has been a push to create better bibliometric tools and better citation search engines and recommendations systems (e.g. [6]), there instead of finding the most relevant papers, now brought out of the background, is the problem of what to do with the papers given the researcher cannot read and perform the level of requisite critical thinking and analysis that is needed on all or even

K-CAP2017 Workshops and Tutorials Proceedings, 2017

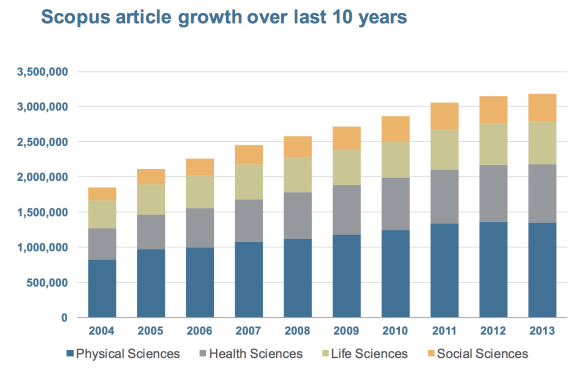

Figure 2: Scopus bibliographic data article growth. From [3]

likely a small percentage of papers produced in a normal literature review process. We believe methods and new human-machine processes are needed to enable the next generation of human-driven scientific analysis, those that go beyond recommending papers to read and instead collaboratively work with human researchers to organize and aggregate findings towards the development and creation of new research directions and experimentation.

3 MORE EFFICIENT RESEARCH CYCLES WITH MACHINE READING

Given for example several many research papers (e.g. 20-40 papers) on a particular variable or construct of interest, averaging between 8-12 pages, the researcher may spend between 2 and 4 days annotating and synthesizing what would amount to a meta-analysis over the set of papers to find the information they need to perform the necessary critical thinking that drives hypothesis formation (taking place in the **predict** step). If instead there were semi-automated processes that, together with the researcher, extracted: variables of interest, relationships, and experimental trends[1], then, some significant amount of time could be saved from, among others, the traditional literature review and analysis tasks that occur during a research cycle; suddenly days of manual annotation and relationship summarization are reduced to minutes or hours. This is in fact an area where both macro- and micro-reading techniques can play a significant role. During macro-reading activities, a collection of research articles are skimmed to extract broad phenomena like variables or methods used in specific articles (e.g. [7, 10]) while micro-reading activities are focused on specific passages of the scientific articles to extract hypotheses and result interpretations (e.g. [9, 11]). These results are used to automatically generate both structured representations of scientific findings and human-readable natural language reports.

4 CONCLUSIONS

The amount of scientific data being generated is growing at a faster rate every year and human ability to continue to sufficiently include and reason over these vast amounts of knowledge is already being challenged. Gone are the days where research in sub-disciples grew at a slow and steady rate and where researchers and their graduate students could adequately review and synthesize findings as they build on prior works. And whereas some would say that the amount of data being generated bids a farewell to traditional scientific methods and processes [1] we take an opposing view and argue that it is not the process or methods but the accessibility of the results to our analysis tools that impedes new rates of progress; we see the incorporation of machine reading research and methods (along with work from other and related fields [12] i.e. research on the scientific process itself [2]) to introduce structure over the scientific finding disclosure process, still largely in unstructured natural language text, as a useful means to enable more efficient and indeed, next generational, science.

ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA).

REFERENCES

[1] Chris Anderson. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine* 16, 7 (2008), 16–07.
[2] Kwame Asante, Eric Barbour, Lauren Barker, Melanie Benjamin, Sara D Bowman, Andrew P Boughton, Erin Braswell, Chelsea Chandler, Nan Chen, Sam Chrisinger, and et al. 2017. Open Science Framework. (May 2017). osf.io/4znzp
[3] Elizabeth Dyas. 2014. Scopus, Science Direct, and Mendeley. (2014). https://www.slideshare.net/nulibrary/scopus-sciencedirect-and-mendeley Presentation.
[4] Daniele Fanelli, Rodrigo Costas, and John P. A. Ioannidis. 2017. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences* 114, 14 (2017), 3714–3719. https://doi.org/10.1073/pnas.1618569114 arXiv:http://www.pnas.org/content/114/14/3714.full.pdf
[5] Timo Hannay. 2015. Science's Big Data Problem. (Aug 2015). https://www.wired.com/insights/2014/08/sciences-big-data-problem/
[6] Gary King, Patrick Lam, and Margaret E Roberts. 2017. Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science* (2017).
[7] Tom M Mitchell, Justin Betteridge, Andrew Carlson, Estevam Hruschka, and Richard Wang. 2009. Populating the semantic web by macro-reading internet text. In *International Semantic Web Conference*. Springer, 998–1002.
[8] Mai T Pham, Lisa Waddell, Andrijana Rajić, Jan M Sargeant, Andrew Papadopoulos, and Scott A McEwen. 2016. Implications of applying methodological shortcuts to expedite systematic reviews: three case studies using systematic reviews from agri-food public health. *Research synthesis methods* 7, 4 (2016), 433–446.
[9] Chris Quirk and Hoifung Poon. 2016. Distant Supervision for Relation Extraction beyond the Sentence Boundary. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 1, Long Papers (2016).
[10] Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2016. Neural Architectures for Fine-grained Entity Type Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 1, Long Papers (2016), 1271–1280.
[11] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 455–465.
[12] Anna Elisabeth van 't Veer and Roger Giner-Sorolla. 2016. Pre-registration in social psychology – A discussion and suggested template. *Journal of Experimental Social Psychology* 67, Supplement C (2016), 2 – 12. https://doi.org/10.1016/j.jesp.2016.03.004

[1] We see here a need for the continued work related to design and development of scientific research registrations processes and conceptual taxonomies (see e.g. [2, 12])

[2] E.g. Towards taxonomy development for appropriately labeling scientific concepts and relationships