Some questions for information science arising from the history and philosophy of science*

Henry Small

SciTech Strategies Inc., Bala Cynwyd, PA 19004, USA
hsmall@mapofscience.com

The occasion of the BIR conference has prompted me to say a few words about research possibilities in information science that are suggested by my background in the history and philosophy of science. I do not represent the constructivist viewpoint and am more traditional in my belief that science is or should be evidence-based and is not predominantly an interest-driven and socially constructed activity.

The first research problem information scientists might address is the nature of scientific discovery. We know that from an information perspective, discoveries are often associated with high citation rates after the fact, and that many discoveries are not recognized as such for many years, although some enjoy immediate recognition. We do not have a clear understanding of the reason for this differential. Nor do we understand how major discoveries can be distinguished from more modest but important research findings and advances. From a prospective point of view, it has been argued that discoveries are novel associations of facts or ideas that had not been previously connected. For example, Don Swanson [5] proposed the idea of "undiscovered public knowledge" where we connect different existing bodies of knowledge which, to some extent, can be anticipated by finding indirect pathways through the knowledge network. However, many discoveries involve novel or unanticipated entities or mechanisms. For example, the hypothesis that CRISPR was a bacterial defense mechanism against invading viruses was initially arrived at by comparing CRISPR "spacer" sequences against viral gene libraries and thus was an inductive process [3]. Other discoveries are more deductive in nature, for example, predicting some known empirical result from theory.

Another type of "discovery" that needs attention is the invention of new methods. The importance of methods in contemporary science is revealed by an analysis of the most cited papers in almost any field of science. It might be argued that methods are now driving science. We know very little about how new methods are invented or how old ones enhanced. Do methods emerge from basic research or do they represent a separate evolutionary path more akin to technological developments? Finally, we are interested in the applications of methods in the conduct of basic research. Obviously, they are a source of evidence to test theories, but also data is collected for the sake of collecting and stored

* A companion video is hosted at https://youtu.be/xOpFBOr0WPg.

in computer databases. Another question is whether theories rely on existing methods for testing or require the development of new methods?

If we look closely at some historical cases, for example, the discovery of the neutron in 1931, we see that scientists had initial inklings prior to the discovery that an electrically neutral massive particle existed. This takes us to the next historical process that needs more research, namely confirmation [4]. How are scientific hypotheses, theories, or hunches confirmed or corroborated? Is Bayes's rule sufficient to account for most historical cases? If so, can the "prior" and condition probabilities required by the Bayesian approach be derived from historical records or statements of scientists? Can informetric and text-based methods be devised to identify competing or alternative theories? Or are there alternatives to a probabilistic theory of confirmation such as consilience or coherence of a knowledge network?

Implicit in my discussion of the problems discussed above are the application of methods for clustering and mapping scientific communities and the ability to delineate structures of leading ideas and concepts at the specialty level [1]. Fortunately, very effective computation methods have been devised for detecting community structures using bibliographic databases most notably citation indexes. More research is needed, however, into studying how these structures evolve over time. Do research areas go through a lifecycle and how do we identify their starting and ending points? Where do discoveries and methods fit into the cycle and can we find evidence of confirmation or disconfirmation occurring? In what sense do these clusters or communities define what Thomas Kuhn called "paradigms" [2]? This construct has remained elusive and undefined. Kuhn also proposed that science undergoes periodic major upheavals called revolutions. We should be able to detect such events, even if only retrospectively, given adequate historical datasets, by studying changes in terminology or cited references. He also proposed that we should see micro-revolutions at the level of small scientific communities. Is there any evidence that micro-revolutions are occurring and how do they differ from their larger brethren? This will of course require taking a deep dive into specific specialty communities, for example covid-19 or CRISPR, which is an approach not currently favored in the informetrics community. In my view without detailed longitudinal case studies, guided by large scale clustering or community detection analyses, we will not be able to get to the bottom of these questions.

Finally, I should mention various approaches to analysis of scientific texts, and particularly citation context analyses. I do not think that the analysis of negative sentiments in citation contexts will be that fruitful because scientists are reluctant to engage in public criticism of their colleagues in print. More productive are studies of the degree of concept uncertainty as indicated by hedging. However, much broader in scope is the use of what might be called "epistemic labeling" in scientific contexts. If we look, for example, at highly cited papers we find not only consensus in the citation contexts on the meaning of cited texts, but also consistency in their labeling as "discoveries", "advances", "methods", "reviews", "databases", etc. and the use of other terms that indicate the cited

concepts' epistemic role, which are often expressed in relational terms, such as "causing", "explaining", "predicting", "confirming", etc. Linguistic methods will be required to delineate, for example, what is "explained" by what, or what is "caused" by what. Obviously, such vocabulary analyses dovetail with some of the questions raised above. Whether the study of such epistemic terms and relationships will take us closer to understanding the "logic of science", as was long the objective of empiricist philosophers of science, remains to be seen.

References

1. Klavans, R., Boyack, K.W.: Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? Journal of the Association for Information Science and Technology 68(4), 984–998 (2017), doi:10.1002/asi.23734
2. Kuhn, T.S.: The structure of scientific revolutions. The University of Chicago Press, Chicago, 2nd edn. (1970)
3. Lander, E.S.: The heroes of CRISPR. Cell 164(1–2), 18–28 (2016), doi:10.1016/j.cell.2015.12.041
4. Small, H.: Past as prologue: Approaches to the study of confirmation in science. Quantitative Science Studies (2020), in press, preprint available at https://bit.ly/Small2020QSS
5. Swanson, D.R.: Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science 38(4), 228–233 (1987), doi:dq8phz