Rich Context

The social sciences are at a crossroads. The enormous growth of the scientific enterprise, coupled with rapid technological progress, has created opportunities to conduct research at a scale that would have been almost unimaginable a generation or two ago. The rise of cheap computing, connected mobile devices, and social networks with global reach allows researchers to rapidly acquire massive, rich datasets; to routinely fit statistical models that would once have seemed intractably complex; and to probe the way that people think, feel, behave, and interact with one another in ever more naturalistic, fine-grained ways. Yet much of the core infrastructure is manual and ad-hoc in nature, threatening the legitimacy and utility of social science research.

We can and must do better. The great challenges of our time are human in nature - terrorism, climate change, the use of natural resources, and the nature of work - and require robust social science to understand the sources and consequences. Yet the lack of reproducibility and replicability evident in many fields(*1–5*) is even more acute in the study of human behavior both because of the difficulty of sharing confidential data and because of the lack of scientific infrastructure. The central argument we advance in this monograph is that advances in technology—and particularly, in automation—can now change the way in which social science is done. Social scientists have eagerly adopted new technologies in virtually every area of social science research—from literature searches to data storage to statistical analysis to dissemination of results.

A major challenge is search and discovery. The vast majority of social science data and outputs cannot be easily discovered by other researchers even when nominally deposited in the public domain. A new generation of automated search tools could help researchers discover how data are being used, in what research fields, with what methods, with what code and with what findings. And automation can be used to reward researchers who validate the results and contribute additional information about use, fields, methods, code, and findings.(*6*)

In sum, the use of data depends critically on knowing how it has been produced and used before: the required elements what do the data ***measure****,* what ***research*** has been done by what ***researchers,*** with what ***code***, and with what ***results***.   Acquiring that knowledge has historically been manual and inadequate. The challenge is particularly acute in the case of confidential data on human subjects, since it is impossible to provide fully open access to the source files.

This monograph provides pathbreaking contributions of different approaches to automating the collection and codification of knowledge from publications and people.   Each paper summarizes technological approaches to applying text analysis techniques on a series of different publication corpora to identify the datasets referenced in each publication and draw out the required elements.   The authors have been identified as a result of an international competition that challenged computer scientists to find ways of automating the discovery of research datasets, fields & methods behind social science research publications.. They are

1. GESIS : Wolfgang Otto, Katarina Boland, Dimitar Dimitrov, Behnam Ghavimi, Narges Tavakolpoursaleh, Andrea Zielinski, Karam Abdulahhad
2. KAIST: Haritz Puerto San Roman, Hong Giwon, Cao Minh Son
3. Paderborn University: Rricha Jalota, Nikit Srivastava, Daniel Vollmers, René Speck, Michael Röder,Dr. Ricardo Usbeck, Prof. Dr. Axel-Cyrille, Ngonga Ngomo
4. Allen AI: Waleed Ammar, Christine Betts, Daniel King, Iz Beltagy

We also will have a contribution from Daniel Acuna, Syracuse University

We envision additional contributions from the technical judges as well as the social science judges.

1. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018).
2. A. Dafoe, Science deserves better: the imperative to share complete replication files. *PS Polit. Sci. Polit.* **47**, 60–66 (2014).
3. J. P. A. Ioannidis, Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
4. N. Young, J. Ioannidis, O. Al-Ubaydli, Why Current Publication Practices May Distort Science. *PLoS Med* (2008).
5. G. Christensen, E. Miguel, Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **56**, 920–980 (2018).
6. T. Yarkoni *et al.*, “Enhancing and accelerating social science via automation: Challenges and Opportunities” (2019).