When data science meets social sciences: the benefits of the data revolution are clear but careful reflection is needed

blogs.lse.ac.uk/impactofsocialsciences/2017/03/01/when-data-science-meets-social-sciences-the-benefits-of-the-data-revolution-ar

3/1/2017

*Contemporary social sciences unquestionably benefit from the growing accessibility and availability of data sources, and the impressive developments in computational tools for data collection and analysis. However, **Marta Stelmaszak** and **Philipp Hukal** emphasise the importance of continued careful reflection when using new forms of data and methods. Any analysis of data requires reflection on the agency that went into defining, recording, and disseminating pieces of information; the fact that computational tools are increasingly deployed to conduct social research must not challenge the foundations upon which social inquiry is based.*

This post is part of our digital methodologies series, arising from a stream at the ISA Logic and Methodologies conference, September 2016, organised by Mark Carrigan, Christian Bokhove, Sarah Lewthwaite and Richard Wiggins.

Ever-larger parts of everyday social activity are influenced by vast, pervasive systems of digital computing technology. Such information systems produce digital information continuously, simultaneously and unequivocally from any interaction they afford with the user: be it the control of home appliances, tracking of fitness activity, or studying a course online. Data collected from these systems is used in research studies across the social sciences and beyond. Yet, only a minor share of scholarly work is concerned with reflecting on the mechanisms that generate, process, and disseminate digital information. The few existing studies reveal a worrying discrepancy among disciplines. Data science, for one, focuses on the practical aspects of efficiently creating, collecting, and handing digital information, often with limited concern about what governs these activities. Meanwhile, social sciences may be wary of the attendant pitfalls of new data sources and methods, but must also maintain rigour and relevancy of their work.

In this context, following the tradition of its annual Social Study of IT workshop, the LSE invited submissions to its Open Research Forum. With the title 'Data science meets social sciences: understanding the origins, means and consequences of social computing', the event gathered a dozen junior researchers and an audience of more than 50 to map a way forward for contemporary social research that uses computational tools and realises the potential of new and large datasets. Here, the organising committee offers a synthesis of the day and the key takeaways from discussion among PhD students, early career researchers, and senior academics from all over Europe and across the social sciences. Some common themes emerged.
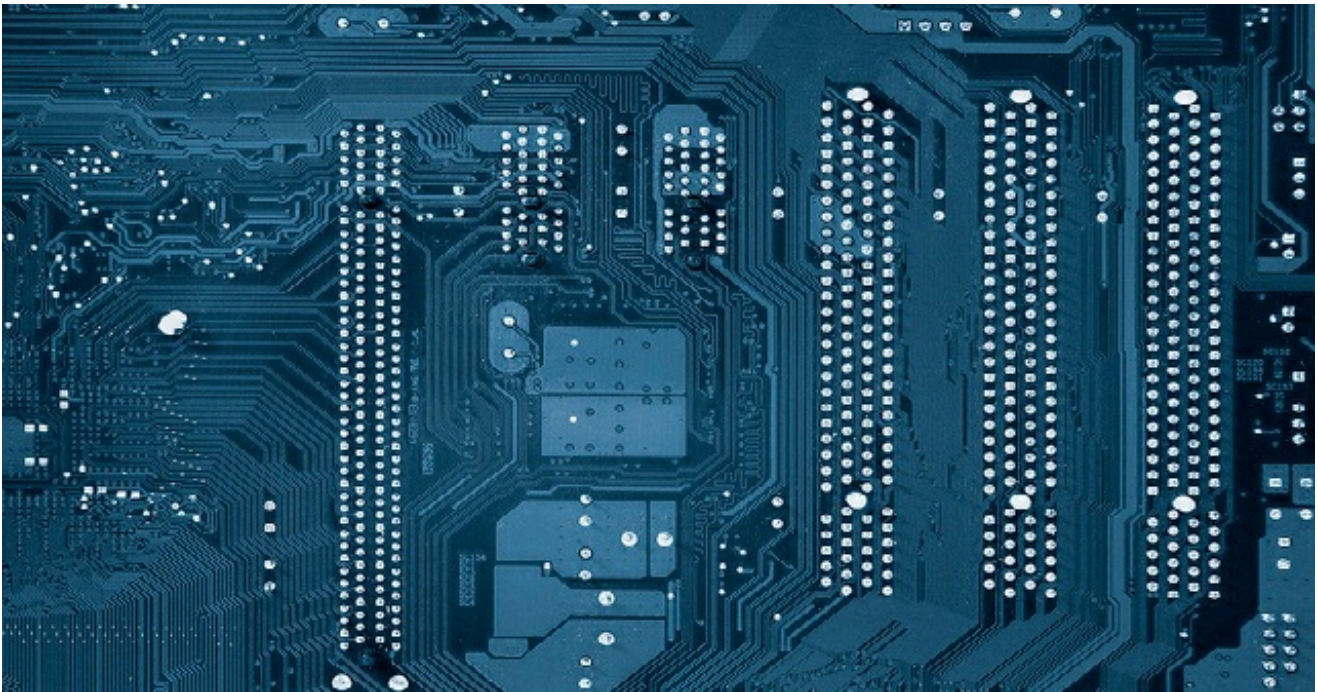
Image credit: TextureX Motherboard Circut blue stock photo Tech Texture by Texture X. **This work is licensed under a CC BY 2.0 license.**

Social sciences benefiting from the data revolution

Firstly, while acknowledging the great promise and opportunities of novel data sources and tools, the workshop also urged vigilance, with researchers warned of the pitfalls caused by the lack of mindful approaches or by scientific discourse being guided by fads. Only when social science researchers remain committed to the virtues of their craft can the mooted benefits of the data revolution deliver on their promise to strengthen social research.

And the social sciences do undoubtedly benefit from the data revolution. The growing accessibility and availability of data sources, in conjunction with impressive developments in computational tools for data collection and analysis, inevitably advance social science research. Many novel and promising data sources are available for scholarly work, with minimal effort required on the part of the researcher to collect the data; be it a simple download from sites such as data.gov.uk, or the utilisation of widespread interface capabilities such as APIs . These sources arguably allow researchers to capture phenomena that were previously either unobservable or even non-existent.

Consider a platform such as GitHub.com, used by software developers all over the world to share projects and collaborate. Storing the software code and its entire editing history of almost 35 million technology artefacts, GitHub's accessible data repositories allow the study of software development at the level of individuals, teams, and entire networks with unprecedented detail and scale. Also, methods that are by no means novel to the social sciences are blossoming again as a function of the accessibility of powerful open-source tools and the plethora of training opportunities available. For instance, methodologies for social sequence analysis have long existed in sociology but are only now gaining attention as a way to complement variance and process-based theories. Similarly, network analysis has long occupied social science researchers, yet a new wave of studies leverages advances in computational techniques for inferences in a degree of detail and complexity that was previously unthinkable. Initiatives like Figshare, which allows researchers to share, use and discuss each other's datasets, indicate a maturation of computational social research as a whole. Such movements point to a much needed push towards the transparency and replicability of studies drawing from novel data sources and methods.

Careful reflection is needed

Despite such glowing promise, calls for careful reflection were common throughout presentations and panels. The reflection spanned three issues in particular:

Contextualisation of digital data in research

All data is social data. Information systems are complex, dynamic compositions of digital technology shaped by social interaction. Whether it's a proprietary business system or a public database holding energy consumption data, the data collection, processing and dissemination through these systems reflects some form of purposeful social agency. Therefore, anyone studying the output of these systems should be aware of the mechanisms that generated and manipulated the data until the point of collection. Objectivity and neutrality are often assumed when researchers tap into the myriad newly accessible data sources. In fact, careful analysis of data requires reflection on the agency that went into defining, recording, and disseminating pieces of information. Through reflection, the fallacy of taking system-generated data as objective fact can be mitigated.

Further training across social sciences

Advanced methods demand advanced training. Methods are maturing and there are currently many tools available to understand social data. In order for researchers to sort the wheat from the chaff, sufficient training efforts must be made. Whether for roles as authors, reviewers, or teachers, the training of social science researchers must complement traditional method and philosophy training with computational techniques. Many universities offer training as part of their coursework. Outside of academic curricula, a variety of offers in the form of massive open online courses (MOOCs) are available (e.g. edX or FutureLearn) as are interactive training programmes such as DataCamp. Whatever the exact form, personal preference, or skill level, some degree of literacy in computational analysis is essential to maintain rigour in research.

Focus on the core craft in social enquiry

The contextualisation in ontological and epistemological discussions is essential to retaining rigour in social science work. In his well-received keynote, Professor John L. King urged the audience to avoid hubris; the relationship between data and theory might be adjusted but it will not be dismissed. Prof King's talk reminded the audience that data helps researchers to develop, reflect and question theory. Yet, data does not and will not replace theory. To paraphrase an Andrew Abbot's example: a variable like gender might describe the phenomenon 'difference in pay'; but it does not explain it. As Prof King pointed out, the fact that computational tools are increasingly deployed to conduct social research must not challenge the foundations upon which social inquiry is based. To this time paraphrase Sutton and Staw's seminal discussion of things that theory is not: method is but one source of strong theory. Yet, in social sciences the method is not a contribution in its own right and ontological and epistemological assumptions need to be reflected upon in every research design in order to be meaningful contributions.

Outlook

Overall, the discussions held during the Open Research Forum demonstrate the huge opportunities new forms of data and methods offer social science research. However, in order to maintain rigour and relevance, social scientists must not just adapt to keep pace with these opportunities, but also embrace novel techniques for fundamental discourses in the philosophy of sciences and their quest of social inquiry. In the end, it is this reflection and investigation that makes social science distinct from data science.

LSE Information Systems and Innovation Group is a centre of expertise on information technology (IT) innovation and concomitant organisational and social change. It is one of the largest groups of its kind in the world, and is well known for its research in the social, political and economic dimensions of information and communications technology.

LSE Information Systems and Innovation Group hosts various events under the theme of 'Social Study of ICT'; recurring formats are seminars, workshops, and open research forums for academics of all levels. More details are

available here or you can follow the organisers on twitter: @ssitorf

This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

Marta Stelmaszak *is a Doctoral Researcher in the Information Systems and Innovation Group, Department of Management at the LSE. Her research focuses on data and measurement.*

Philipp Hukal *is a Doctoral Researcher in the Systems and Management Group at Warwick Business School. His research focuses on digital platforms and digital innovation.*