Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December: 1–6

DOI: 10.1177/20539517211040198
journals.sagepub.com/home/bds
SAGE

A view from data science

Anna Sapienza[1] **and Sune Lehmann**[2]

Abstract

For better and worse, our world has been transformed by Big Data. To understand digital traces generated by individuals, we need to design multidisciplinary approaches that combine social and data science. Data and social scientists face the challenge of effectively building upon each other's approaches to overcome the limitations inherent in each side. Here, we offer a "data science perspective" on the challenges that arise when working to establish this interdisciplinary environment. We discuss how we perceive the differences and commonalities of the questions we ask to understand digital behaviors (including how we answer them), and how our methods may complement each other. Finally, we describe what a path toward common ground between these fields looks like when viewed from data science.

Keywords

Interdisciplinary approach, social data science, digital behavior

This article is a part of special theme on Machine Anthropology. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/machineanthropology

Introduction

A slice of big data

Within a very compressed time period, the world has been transformed by Big Data. Technological advances have allowed the aggregation of enormous stores of information describing our digital behavior (Lazer et al.,2009), but with big data also other elements come into play: variety, velocity, and value (Dumbill,2012; Gantz and Reinsel,2011). Here, we discuss how differences in methodologies can impact the way these forms of data shape new understanding of individual behaviors and societies.

Getting to the transdisciplinary future

Behavioral data present an opportunity to build rich models of individuals and answer new research questions which were previously unanswerable because we simply did not have data. To draw the maximal benefit from new data sources, therefore, we need to invent new questions alongside new methods. We need to build transdisciplinary approaches merging social science (SS) and data science (DS) instead of complementing them (Blok and Pedersen,2014). Until then, data and social scientists face the challenge of effectively building upon each other's approaches to overcome the limitations inherent to each field and capture the full potential of big data.

Long collaborations

In our experience working in an interdisciplinary environment,[1] a key step is fostering long-lasting collaborations. Trust and patience are important to connect the expertise of fields that often look down on one another and care passionately about very different questions. It takes time to establish a common ground of shared questions and scientific values on which this collaboration builds.

When envisioning social data science (SDS), data and social scientists share the goal of developing a field that

[1]Copenhagen Center for Social Data Science, University of Copenhagen, Denmark
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

Corresponding author:
Anna Sapienza, Copenhagen Center for Social Data Science, University of Copenhagen. Nørregade 10, 1165 Copenhagen, Denmark.
Email: ansa@sodas.ku.dk

uses vast and complex data to study social and digital behaviors in situ and design models capable of inferring patterns from that data. Patterns that we need to interpret through the lens of social theories, by considering behavior as part of the cultural environment where it occurs (Shah et al.,2015).

The devil is in the details

We believe that part of the difficulty in building partnerships is related to which questions each field finds interesting. The first challenge, therefore, is to find concrete questions that are relevant to and publishable within different fields.

Here, we offer a DS perspective on the challenges arising when trying to establish the interdisciplinary field of SDS, by discussing: how we experience differences and commonalities of *the questions* we ask; how DS perceives differences *in methods* and how they can potentially complement each other; and paths toward common ground between the fields.

The questions, their answers, and how we talk about them

Disciplinary audiences are the norm

Although many fields share similar goals, specific research questions are different across fields. We do not generate knowledge in an abstract way. We write papers. Papers with disciplinary audiences and standards of evaluation developed within specialized journals. Currently, interdisciplinary research has no common evaluation criteria and often encounters challenges in finding funding (Bromham et al.,2016), getting papers accepted and cited in the short-term (Wang et al.,2017). Given that scientific achievements are mainly measured via factors, such as the *h*-index, these issues are tied to the researcher's career perspective and form a strong incentive to favor disciplinary audiences. Thus, to publish a paper, SDS researchers must choose a disciplinary venue, which implies choosing questions that are relevant to a disciplinary audience. Yet, as a rule, computer science questions (e.g., developing powerful algorithms, performance benchmarks, etc.), are not interesting to SS journals and vice versa. This is a key problem for moving forward.

Finding good questions as data scientists

DS is not quite computer science, so let us be a bit more specific about how we think about identifying questions. As social scientists, data scientists are interested in describing behaviors. But what are the differences? As data scientists, we start from large datasets and search for questions in much the same way that natural scientists ask questions about nature. Our criterion is that findings must be surprising/interesting/non-trivial, but yet supported in a rigorous statistical sense by the data we are studying.

The secret trick (we believe) is that alongside the questions, we search for the *variables* defining the questions. As pointed out by Milner (2018), identifying the right variables is essential to the Natural Sciences. Milner uses Newton as an example. The genius of Newton was not the mathematical theory, but that he (within the entire natural world) identified the variables which connect the behavior of apples falling to the ground with planetary motion: *force* and *momentum*.

Thus, we explore large datasets as someone from the natural sciences explores the physical world. This means that we are not hypothesis-driven in the sense of many SSs (more below). As we explore the data, we form informal theories of what might be driving the patterns we see. We learn what the right questions (and variables) are.

Brockmann et al. (2006) provide an example of this process. Here, the authors departed from the established norms of understanding travel behaviors, by looking at dollar bills movements. They realized that mobility could be understood via two newly defined variables: scale-free jumps and long waiting times between displacements. Identifying these variables opened up new questions regarding the description of travel patterns that had not been explored previously—exploration that is possible without *p*-hacking (Head et al.,2015).

What do SS methods look like to us?

We start with the caveat we understand that there are philosophical and methodological chasms dividing SSs. We cannot address everything here, so we mainly discuss a simplified "hypothesis-driven research" (which is (a) not the norm in all SSs and (b) not exclusive to the SSs). But the hypothesis-driven paradigm (and its deductive nature) presents a useful contrast to the described natural/DS approach.

Hypothesis-driven research tends to appear when things are complicated. Milner uses biology as an example, but we believe it can be extended to some SSs, for example, quantitative research in sociology, psychology, political science, and economics:

> Researchers in those fields study complicated, irreducible systems (living organisms), have limited experimental probes, and are often forced to work with small data sets. Unavoidably, the most common experimental protocol in these fields is to poke at a complex living system by giving it a drug or chemical and then measuring some indirectly related response. Those experiments live and die by the statistical test.
>
> Milner (2018)

A clear example of this experimental design is the practice of *preregistration*, which is becoming predominant

especially in psychology, economics, and political science. When research is preregistered, not only the questions/hypotheses need to be specified *but also the variables* and tests. As exemplified by Milner's quote, this approach is sometimes necessary, but it eventually limits the exploration, especially when large datasets are available.

Another "locking in" of variables and questions can occur when social scientists work in a highly theory-driven way and frame their work in terms of traditional "big questions" (Giles et al.,2011) and paradigms (Burrell and Morgan,2017). In sociology, examples of big questions are as follows[2]:

- To what extent is the individual shaped by society?
- To what extent does our social class background, gender, or ethnicity affect our life chances?
- What is the role of institutions in society?

When using theory and hypothesis-driven approaches, finding the data to answer a specific question becomes central. But predefined questions frame the variables, thus, limiting the way information is extracted from large datasets.

Talking about findings

Another cultural difference relates to how we talk about findings. As data scientists, an important criterion for quality and impact is *generalizability* (Lee and Baskerville,2003). As any research on human beings, in one sense, DS findings are highly specific to the dataset under study. Nevertheless—perhaps because we are often revealing new variables to measure and care about—data scientists often declare findings to be highly general. To give a concrete example, the following paper is entitled "Fundamental structures of dynamic social networks" (Sekara et al.,2016). A more general title than "How We Tweet About Coronavirus, and Why: A Computational Anthropological Mapping of Political Attention on Danish Twitter during the COVID-19 Pandemic" (Breslin et al.,2020), even though the data used in both works is based on a subset of the Danish population. We have learned, through our collaborations with social scientists, that this tendency can be an annoyance to some.

Are data scientists then bullshitters?

Despite this difference in communication, we argue that data scientists do not just over-hype their results.

Firstly, DS brings to the table the capability to explore very different, larger, and general datasets.[3] Secondly, applying methods from the natural sciences to social data, can (e.g., through the "discovery" of new, interesting variables) identify new patterns to investigate across settings.

Let us take (Brockmann et al.,2006) as an example again. The authors started by studying bank notes dispersal through about 1M reports from a bill-tracking website. They introduced the concept of displacement distributions, which had previously not been studied. To show that this concept was interesting beyond their initial dataset, they extended their findings by exploring two additional datasets.

Demands for immediate usefulness

Generalizability, however, does not imply the results to have a practical and immediate application in real life. Many SS outcomes (e.g., in economics) are subject to the constraint of being immediately useful. When social scientists are asked to inform policymakers, defining one's own variables of interest is not an option. We have experienced that working on these problems tends to shape how researchers approach science altogether.[4]

Since DS researchers work to identify interesting patterns, our "generalizable" outcomes can turn out to be useless. They do, however, have the benefit that if proven useful, they can be powerful. Sometimes unexpected applications are discovered long after the exploration. Nowhere is this more evident than with pure mathematics, where (initially useless) results from number theory have later formed the basis of ubiquitous cryptographic protocols.

A trade-off

One can imagine a trade-off between research strategy and results' immediate impact. We show this idea in Figure 1. On the one side, we find fields such as physics and mathematics, whose questions/variables are not predefined and whose results generalize to the entire known universe. However, many results have no practical value, and when they do, findings sometimes need decades to find an application. We provided the example of number theory but there are many others: the theory of conic sections had no practical purpose until Kepler found that they describe planetary orbits; non-Euclidean geometry, formulated by Riemann (among others), did not find practical use until Einstein used it to model the universe, etc.

On the other side, we locate areas whose approach, for the reasons mentioned above (e.g., demands for highly specific answers, complexity combined with limited data, theory-driven approaches, etc.), entails a more rigorous definition of both questions and variables—parts of SS, biology, and engineering. This connects to the idea of the "purity" of sciences depicted in Figure 2.

We think that DS currently exists somewhere between these extremes: we choose the variables and frame our questions in relation to them. However, this freedom comes at the cost of reducing the immediate practical usefulness of our results.

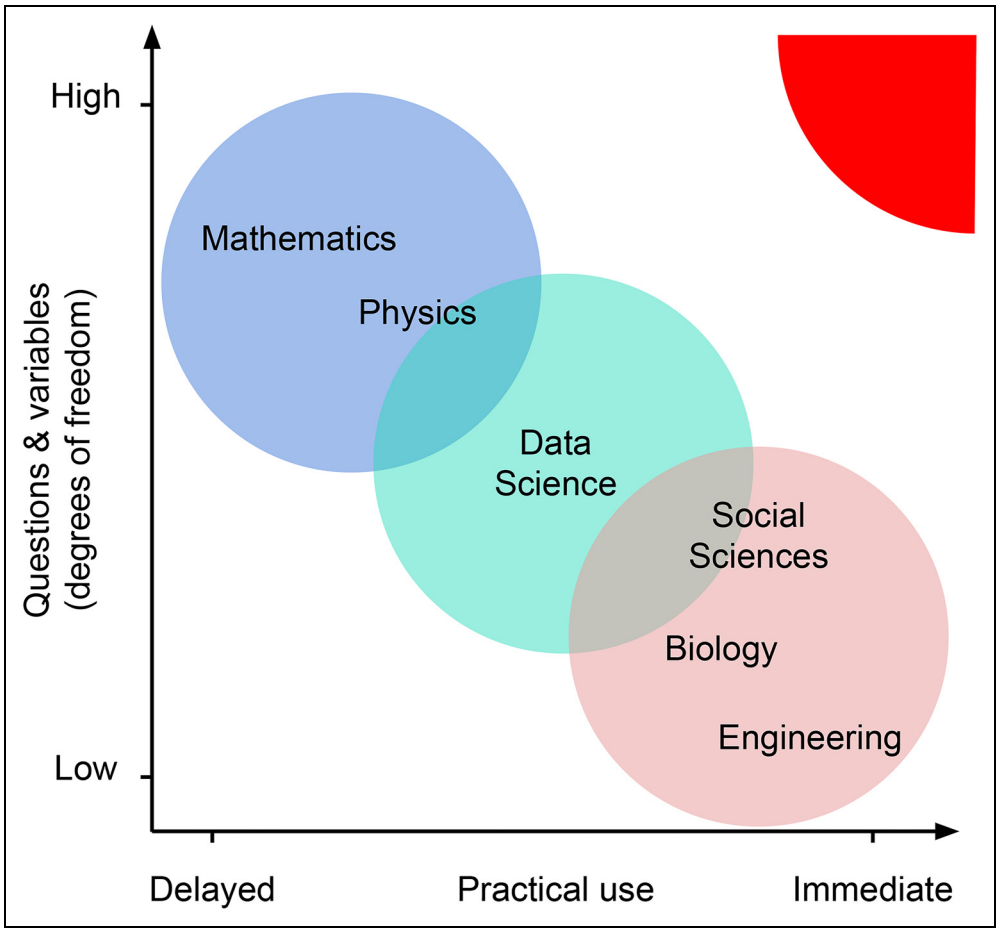

Figure 1. Trade-off between the research approach and the practical use of findings.

Ideally, the goal of any research field is to understand the world completely. To have full availability of fantastic data, freedom to identify the variables to describe the world; to design models and theories that are both general and of immediate practical use. To hold the enviable positioning at the top-right corner of Figure 1 (in red). This sketch is not a deeply considered view of all science, but rather a jumping-off point for thinking about how now data and the combination of SS and DS practices can provide new chances to optimize the trade-off.

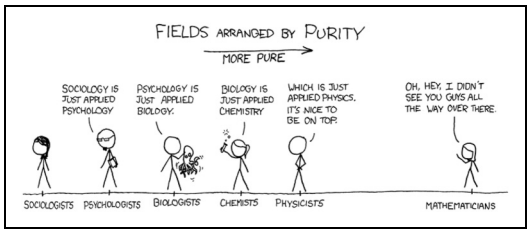

Figure 2. The comic "Purity" (https://xkcd.com/435/). CC Non-Commercial 2.5 License.

Wrapping up

Moving forward

We have attempted to describe some of the problems that arise when collaborating within SDS: How—due to different traditions and audiences—data and social scientists are trained to provide different styles of answers when designing their research and writing papers. Now, we describe how we see the road going forward.

Shared questions and methods

We believe that the road ahead lies in collaborating to combine methods at scale to find deeper, and more general answers from high-resolution data. Perhaps data will allow to identify new and powerful variables for SS. To understand these data, we need to combine data sources providing both the behavior and the perception of it. Moreover, we need to rely on the expertise of data scientists in identifying the important variables. And we need social scientists to providing a theoretical explanation driving the observed behavioral patterns. This combination

of data and approaches will also help reframe our questions into a shared view involving *how* and *why* different social behaviors occur. Neither field can meet these challenges alone.

Building bridges

It is not just within a methodology that we can find common ground. We started this essay by noting that until we form new disciplines, collaborations need time. We have experienced that, over time, perhaps we can also find new and shared questions and answers. To collaborate across disciplines, beyond combining different methods, we need to know each other well enough to discuss freely. We need to understand each other's worlds well enough to get over the existing gap between approaches and look for variables, questions, datasets, and answers that everyone finds interesting. There are many ways this can be done. One we have found is to simply bring someone along for the paper. We have experienced brilliant and creative ideas from including an anthropologist in the data exploration of a typical DS project or having a data scientist join a group of social scientists analyzing electronic questionnaires and free text data. More generally, we have to consider the research process itself. It is not only the differences in questions and approaches, but also the differences in scientific practices. We think that practical experience is a good way of building this type of experience.

Interdisciplinarity is necessary

As we approach a transdisciplinary future, we need to make it through the first wave of interdisciplinary collaborations.

To understand ever-growing data streams, we genuinely believe that a new breed of transdisciplinary researchers is necessary. "[T]he large datasets that describe the inner workings of complex social systems […] cannot be adequately understood from a single disciplinary perspective" (Chang et al.,2014). The amount of data, currently available to investigate digital behaviors, cannot be analyzed without DS and machine learning methods and cannot be fully understood without social theories linking the motivation behind actions to the observed behavior.

To get there, we must already now start creating an infrastructure that rewards collaboration across fields via new journals and funding schemes. We need editorial boards and funding bodies with interdisciplinary backgrounds—who recognize that interdisciplinarity is difficult and takes time. Now is the time to set the stage for how and where this transdisciplinary research will be performed and published.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Paper developed under DISTRACT—The Political Economy of Distraction in Digitized Denmark, funded by the ERC Advanced Grant—Horizon 2020 (834540).

ORCID iDs

Anna Sapienza https://orcid.org/0000-0002-0842-7987
Sune Lehmann https://orcid.org/0000-0001-6099-2345

Notes

1. At the Center for Social Data Science—University of Copenhagen.
2. Adapted from lhttps://revisesociology.com/2017/08/08/big-questions-a-level-sociology/.
3. Large-scale datasets have many shortcomings relative to traditional SS data-types as questionnaires. For a discussion of these topics, we recommend (Salganik,2019).
4. SSs are not a coherent single area of science. Similarly, a strong focus on usefulness is present in engineering, thus there is more nuance to this story than we have room for here.

References

Blok A and Pedersen MA (2014) Complementary social science? Quali-quantitative experiments in a big data world. *Big Data & Society* 1(2): 2053951714543908.

Breslin SD, Enggaard TR, Blok A, et al. (2020) How we tweet about coronavirus, and why: A computational anthropological mapping of political attention on Danish Twitter during the covid-19 pandemic. In: *The COVID-19 Forum III. University of St. Andrews*, volume 18.

Brockmann D, Hufnagel L and Geisel T (2006) The scaling laws of human travel. *Nature* 439(7075): 462–465.

Bromham L, Dinnage R and Hua X (2016) Interdisciplinary research has consistently lower funding success. *Nature* 534(7609): 684–687.

Burrell G and Morgan G (2017) *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Abingdon: Routledge.

Chang RM, Kauffman R J and Kwon Y (2014) Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems* 63: 67–80.

Dumbill E (2012) *Planning for Big Data*. Sebastopol: O'Reilly.

Gantz J and Reinsel D (2011) Extracting value from chaos. *IDC iview* 1142(2011): 1–12.

Giles J et al. (2011) Social science lines up its biggest challenges. *Nature* 470(7332): 18–9.

Head ML, Holman L, Lanfear R, et al. (2015) The extent and consequences of p-hacking in science. *PLoS Biology* 13(3): e1002106.

Lazer D, Pentland AS, Adamic L, et al. et al. (2009) Life in the network: The coming age of computational social science. *Science (NY)* 323(5915): 721.

Lee AS and Baskerville R L (2003) Generalizing generalizability in information systems research. *Information Systems Research* 14(3): 221–243.

Milner S (2018) Newton didn't frame hypotheses. why should we?. *Physics Today*.

Salgnaik MJ (2019) *Bit by bit: Social research in the digital age*. Oxfordshire: Princeton University Press.

Sekara V, Stopczynski A and Lehmann S (2016) Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences of the United States of America* 113(36): 9977–9982.

Shah DV, Cappella JN and Neuman WR (2015) Big data, digital media, and computational social science: Possibilities and perils. *The Annals of the American Academy of Political and Social Science* 659(1): 6–13.

Wang J, Veugelers R and Stephan P (2017) Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy* 46(8): 1416–1436.