Schweitzer and Vespignani *EPJ Data Science* 2012, **1**:1
http://www.epjdatascience.com/content/1/1/1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

EDITORIAL Open Access

Editorial

F Schweitzer[1*] and A Vespignani[2,3*]

* Correspondence: fschweitzer@ethz.ch; a.vespignani@neu.edu
[1]ETH Zürich, Kreuzplatz 5, Zürich, 8032, Switzerland
[2]Northeastern University, 360 Huntington Ave., Boston, Massachusetts 02115 USA
Full list of author information is available at the end of the article

Science is changing, all the time. Some consider external factors to be the major driving force, such as new discoveries, or new technologies that provide deeper insights into known systems. Others consider internal factors to be more important, such as the never-ending search for eternal truths or, at least, for the theoretical unification of diverse observations. Yet others believe that science is a social endeavor and its evolution is governed by forces very similar to those that cause fashions to change, with new research topics seeming to catch our attention and established topics falling out of vogue. We have come to believe that all of these arguments apply, and that the various impacts on scientific progress are even mutually dependent. Fashionable topics co-evolve with technological advances, as do new discoveries with overarching theories.

However, since about the turn of the century we are facing a new kind of (r)evolution in science that is driven not primarily by new discoveries, theories, or fashions - but by the unprecedented availability of data, which is "out there" and has yet to be explored. The spectacular advances in the natural sciences over the past hundred years have been built in significant part on the capacity to collect and analyze data. So what is the difference now? Most of the time, finding and generating data was a scientific enterprise by itself, tightly knit to engineering and technological developments - from the telescope to the microscope, from the particle accelerator to the mainframe computer. But these advances were basically driven by hypotheses - assumptions about the real world that needed to be tested. Experiments, observations and computer simulations were grounded in scientific reflection about what should be explained in a rational manner, i.e., they largely followed the underlying scientific hypotheses. This is what has changed now. Every 1.2 years, more human-driven socioeconomic data is produced than during all preceding human history combined [1]. Data that is just the outcome of our daily activities and not the result of some well-planned scientific experiments. Yet this data conceals the patterns and regularities at the basis of the understanding of socio-technical and socioeconomic systems.

From an abstract point of view, we can consider the internet and its corollary mobile ecosystem as a gigantic social observatory that records basically every online interaction of human beings, their search queries, their personal preferences, with whom they interact, and what's more, does so in real time. The internet was not set up for that purpose, yet it constantly generates a massive data stream that is just waiting to be analyzed - and even more so, to be understood regarding what it means to be "human" in the age of pervasive communication. The internet is by no means the only source where massive data is generated: mobile phones, travel data, credit card transactions, voting records - other personal data compliments our digital footprints and

could lead to a completely revised image of humans, if aggregated and merged in the right way.

Notably, ICT technologies also permit novel paradigms for data generation that exploit socially intelligent technologies and the wisdom of the crowds. The unprecedented possibilities of "enhancing the human image" with empirical data have already been recognized - not only by social scientists, who want to learn about the way humans interact and influence each other - but also by engineers, computer scientists, physicists, and health scientists (to mention but a few), seeking novel approaches to modeling, forecasting and policymaking in socio-technical systems.

So, where is the challenge? Unexpectedly, in understanding the difference - or the convergence - of science and engineering. The scientific community has not asked for this wealth of data it is now confronted with and is hardly prepared to pose the scientific questions of what to learn from these new empirical sources - beyond testing some already established paradigms. This conceptual gap, as we call it, opens the door for engineering approaches to take over: instead of understanding the generative forces underlying a phenomenon, the focus is on control, on application, on solving problems. There is nothing wrong with this - except if we take it as a substitute for scientific understanding. Do we really believe that data is the message? Or, do we believe that we are still confronted with the problem of identifying the message?

In 2008 the editor-in-chief of *Wired*, a popular magazine with an eye on the latest trends in technology, stated that "faced with massive data, this approach to science - hypothesize, model, test - is becoming obsolete." What he proposes instead is the "petabyte age: Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different."

The idea to launch a new journal, *EPJ Data Science*, was formed exactly around the challenge of tackling massive amounts of data in a scientific manner, by exploring its engineering but even more so its conceptual challenges. Accordingly, the journal's scope goes well beyond technical issues of gathering data from "sensors" or programming issues of data crawlers. It also goes beyond the classical statistical analysis. Our focus here is on identifying new empirical laws emerging from massive data sets and the "How?" question, i.e. on conceptually new scientific methods for analyzing and synthesizing these laws. We want to recognize the picture that is hidden in these massive data streams, to predict its occurrence in a statistical sense, and to control it. But we also want to go further, to the "Why?" question, by linking these findings to theoretical concepts in a broader sense, to understand their origin and their impact.

Going for the latter implies more than plotting data the right way, fitting curves, or mapping regularities to known dynamics. New concepts need to be established - notably about social systems - that support and contain these findings, models of basic social interactions need to be developed to predict a certain outcome on the system level, and we have to pay attention to those empirical findings that do not nicely fit with established theories. This involves all disciplines from the social sciences, in particular social psychology, to computer science, to engineering algorithms and data

retrieval mechanisms, and statistical physics, in order to understand collective behavior. This is the direction where we want to go with *EPJ Data Science*.

Do we need a new journal to cope with this development?

We do believe that even after decades of arguments about the value of transdisciplinary research, science is still very much fragmented. Moreover, this fragmentation is reinforced every day, making it hard for scientists to publish work that falls between the "disciplinary" cracks, and to earn reputations for taking the bigger challenge.

With the assistance and support of an international editorial board reflecting a broad range of academic backgrounds, we intend to change this by providing a platform that encourages the submission of those contributions which do not easily fit within the more rigid disciplinary boundaries of traditional journals or would particularly benefit from a wider exposure through "Thinking out of the disciplinary box".

Last but not least, we aim to provide an interdisciplinary, top-ranked journal that goes beyond archiving technically sound papers, focusing on articles which contribute significantly to the conceptual and technological progress in the field of data-driven science.

How broadly defined is the scope of the new journal?

While we agree that Data Science indeed is not very specific in terms of the areas the data comes from, we have made clear in the "Aims and Scope" Section of the journal that the focus is on systems where the social component is a key to scientific understanding. In the widest sense, this includes social interaction (including animal societies), economic and financial systems, management and business networks, socio-technical infrastructure, health and environmental systems, the science of science, as well as general risk and crisis scenario forecasting up to and including policy advice. While *a priori* the interest lies in finding new empirical laws, or even more fundamental theories, concerning how any natural or artificial complex systems works, it implies that we do not embark on research areas such as genomics, cosmology, or geophysics - which are data-intensive as well, but already have their specialized publication outlets.

What is the role of modeling in the new journal?

We emphasized above that our ultimate aim is to contribute to the scientific understanding of the system or phenomena. Consequently, this goes beyond a statistical analysis of massive data, and also includes models of the generative dynamics underlying it. Ideally, models bridge the (partly unknown) behavior of the actors and the stylized facts observed empirically. However, the development of general models without a direct relation to data is not deemed adequate for our journal.

What is the role of algorithms and code in the new journal?

The engineering perspective is very important for the advancement of the whole field. Gathering massive data and processing it in a computationally efficient manner is not just an afterthought; it decides about the part of the digital world (and conversely that of the human world) we are able to capture. Thus, we encourage submissions that pay attention to these questions, while not losing the focus on the phenomenon at hand. But we are less focused on technical aspects of manufacturing devices designed to record or process data.

Will data sets used in publications be made available by the journal?

Collecting, categorizing and archiving data is also a science by itself, as is the proper selection of existing or to be established data sets for a well-defined scientific purpose. Accordingly, the journal will introduce a particular type of regular article dealing with best practices of collecting specific data sets. Such papers shall make a sample of both the raw and treated data set available to the public as an electronic attachment to the article and discuss in detail the methods used and assumptions made in collecting and treating the data, and thus what bias or limitations this introduces in view of the further use of such data sets for various applications. In particular authors should mention where and how the full data sets can be accessed. In this way the journal also seeks to contribute to increased transparency in data set archiving and usage, as well as foster the establishment of credited data sources and in this way accelerate the advancement of the field.

We have devoted this new journal to all scientists with a generic interest in complex, human activity-related systems, their "microscopic" interactions and their "macroscopic" patterns - to those who are anxious to uncover the message in the data, and willing to understand its fundamental origin.

Frank Schweitzer, Editor-in-Chief

Alessandro Vespignani, Editor-in-Chief

Zurich/Boston, November 2011

Author details
[1]ETH Zürich, Kreuzplatz 5, Zürich, 8032, Switzerland [2]Northeastern University, 360 Huntington Ave., Boston, Massachusetts 02115 USA [3]Institute for Scientific Interchange Foundation, Via Alassio 11/c, Turin, 10126 Italy

Received: 25 November 2011 Accepted: 17 May 2012 Published: 17 May 2012

Reference

1. Manyika James., *et al* (2011) Big data: The next frontier for innovation, competition, and productivity. The McKinsey Global Institute

doi:10.1140/epjds1
Cite this article as: Schweitzer and Vespignani: **Editorial.** *EPJ Data Science* 2012 **1**:1.