IEEE Intelligent Systems

AI and Web Science

The Role of Data Science in Web Science

Christopher Phethean, Elena Simperl, Thanassis Tiropanis, Ramine Tinati, and Wendy Hall, University of Southampton

The proliferation of digital technologies and their increasingly common connection to the Web give rise to new opportunities for areas of study—in academia, business, and government alike. Web science and data science are two such examples, and although they are distinct, they share several similarities and can complement each other to gain greater insights into a huge range of topics. The Web presents such a variety of research challenges that an equally broad selection of techniques are required to gain insights. Since its emergence, Web science has been establishing a set of scientific methods for studying the Web's social and technological networks. This is enabled by the scale of data now available online. Social networking sites, webserver logs, and user-generated content are just a few examples of how the Web enables the creation of data at a rate that demands new ways of thinking, handling, and reporting findings. In order to exploit the vast amounts of data now accessible online, one emerging discipline that is crucial for Web science to synthesize with others is data science, which is developing approaches for extracting information from data so that it can be used in decision making and problem solving.

Principles

Web science studies the sociotechnical relationship between people and the Web, examining not only how the technology behind the Web facilitates the various applications now running atop it, but also the role of the people using it, how their lives are changed by it, and how they play a vital role in shaping and maintaining the Web for future generations.[1] The technical layer—referring to the technologies behind the Web—makes up only a part of what Web science focuses on. Equally important are two further layers: the social layer, which emphasizes the people and the content they create and share online and how the networks among them facilitate this, and the market forces and policy layer, which relates to the interconnected economic and political factors that shape the Web's evolution. As such, Web science is as much about the networks of people brought together by the technology as it is about the underlying technological network itself.

Web science's interdisciplinary aspect is fundamental to its aims of establishing the Web (not just its underlying technologies) as an object to study. The Web's impact on society, and vice versa, receive equal importance and focus. Sociology, politics, law, economics, and anthropology all provide invaluable contributions to the field and are fundamental in ensuring a holistic and societally beneficial analysis of the Web. Vast amounts of data produced by and stored on the Web reveal previously unobservable human phenomena at a grand scale, allowing new insight into society. This data fuels new research questions and possibilities spanning multiple disciplines and methodological approaches.

One approach to analyzing this data could be to draw on data science, which provides methods for dealing with data of all shapes and sizes and is proving useful for handling the rise of *big data*—large amounts of data that might be structured, semistructured, or unstructured. Big data is produced at such a rate that it requires new management policies and analytic methods, but it offers the chance for new insights that capture phenomena in real or near-real time. The rate at which data is created leads to challenges in storing, managing, and using the data productively. Much of this data is produced thanks to the increasing ubiquity of the Web, digital technologies, and connected devices, and it is the potential of unstructured data that can be highlighted as a primary difference between data science and established disciplines such as statistics.[2] The demand for data science skills stems from research, commerce, and government alike. As with Web science, data science requires significant understanding of the subject area to complement programming and statistical skills to

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

ensure a relevant question is asked and answered. New expertise is required, now not only limited to statistics but covering a range of topics from data curation to data analysis, data reporting, and storytelling.

Both these disciplines have arisen through the prominence and increasing pervasiveness of the Web and digital technologies. They can provide new insights into a rapidly changing world and complement each other. To better portray the areas in which they overlap—along with those where they do not—this article will outline their similarities and differences.

Interdisciplinary Approach

Web science and data science are both interdisciplinary subject areas. For Web science, this aspect is key to the discipline and helps to produce an assessment of both the Web's social and technical networks, going beyond what any one subject can produce. The disciplines involved can include syntheses of the humanities, social sciences, hard sciences, and engineering, and could include an assessment of the underlying technology and how this affords or is affected by a phenomenon occurring on top of it. Indeed, discussions have persisted about where Web science's role lies in the computing discipline because its breadth goes far beyond the traditional focus of the subject.[3]

Data science focuses on a core statistical and analytical approach to data, which forms the basis of the discipline, and is assisted by techniques from computer science and its subdisciplines, such as machine learning and AI. These techniques are further complemented with domain insight into a problem, which helps contextualize and design the study, taking into account (where applicable) certain disciplinary aspects that provide direction for the statistical work. Data scientists will understand "the mathematics, statistics and physics necessary to integrate science algorithms into efficient architectures."[4] Following this, expertise is required to report on the findings in a way that makes sense to any invested parties, and this brings in practices from business and visualization; one of data science's key roles is arguably to help understand business and innovation practices. As such, although both subjects share an interdisciplinary backbone, the nature of this mix is different: data science bases its approach on a key set of techniques focusing on and incorporating statistical data analysis, whereas Web science integrates a mix of approaches from across any discipline. Each way produces results that are grounded in the context of the problem they seek to solve, which is essential in considering the complex digital world that has arisen. Both disciplines therefore offer much in terms of the value of the insights that they produce. Web science is constantly aware of Web protocols and technologies, looking at how developments around these can impact human activity—and vice versa. The scale of the Web's impact in the contemporary world means that this touches on most aspects of modern life. Data science can go further in some regards to produce insights into many industries or disciplines where the data involved has little to do with the Web itself—or the people involved in using and shaping it—and as such is more technology agnostic.[5]

Actionable Insights

Both data science and Web science help produce insights above and beyond what individual, traditional disciplines could provide, and the key factor is that these insights are actionable and can be used to identify social or corporate phenomena, implement new policies, and make business decisions. Some existing subject areas will rightly question whether everything is completely problem-driven and actionable, but the affordances of studying data to this scale can still offer valuable opportunities for complementing existing research practices.

In Web science, studying interaction and communities on the Web facilitates new possibilities for understanding a population's social and cultural norms, upon which new policies can be designed that consider both the Web's social practices and technological affordances. The subject has an established goal associated with it: to ensure that the Web remains beneficial to society, meaning that insights are gained that contribute toward this aim.[5] In data science, the data might not be Web related and the goals could be wider than this particular sociotechnical artifact, but studies will often utilize several Web-based technologies, such as semantically tagged datastores**, i**n order to store and process that data. The insights can then be used to develop new business models, identify gaps in a market, or spot trends or anomalies in anything from agricultural data to security data to healthcare data. Similar to network science, which examines not only the Web but also transportation, biological, and other technological networks, data science produces insights that are actionable across a range of domains. It is clear, however, that many of the techniques and resulting insights gained from data science are applicable for use within Web science research. New policies could be implemented based on large-scale analysis of Web data that combines subject expertise and knowledge of social science with the computational tools of data science. In this example, data science's insights and outputs are equally labelled insights from Web science; similarly, if the domain was astronomy, the data science insights would be astronomical outputs. Data science offers the tools and methods to produce insights, but it is within the context of the discipline—Web science or astronomy, or any other—that the insights are actionable.

Research Methodologies

Given the broad range of disciplines with which Web science seeks to integrate, it is unsurprising that there is a mixed set of research approaches. Mixed methods combining interpretive or constructivist approaches with positivism are common and emphasized as essential,[1] and they help Web scientists understand aspects of the Web such as online human behavior, social network formation, and the spread of viral content.

Alternatively, being more closely tied to certain disciplines, data science often has an empirical element at the heart of its methodology. However, new methods and approaches are required compared to traditional statistical approaches, because the data is more complex and is often being created, accessed, and analyzed in real time to provide instant analysis. Given this data's relevance for assessing social phenomena, business insights, and policy provision, however, it is also necessary to contextualize these results, and accordingly, the research design must have an investigative and interpretive element. Therefore, mixed methods is again a suitable approach, but the way in which the methods are combined and the permanent inclusion of an empirical element demonstrates the differences with Web science.

Data science also draws heavily on open innovation, in which data is made available to external researchers or analysts and new ideas are sought. This increases the chance of novel approaches being implemented to produce insights, leading to a greater opportunity for flexibility in the methodological process.

Research Agendas

Web science has been described as being "focused on how we could things better."[6] This refers to the potential to improve understanding about the Web and to ensure that future developments remain beneficial to human society and that it remains "pro-human."[7] In their "Manifesto for Web Science," Susan Halford and colleagues propose that Web science should go further than simply being "a sociology or a computer science of the web," suggesting that it must be a genuine intersection of disciplines that can examine micro- and macro-phenomena on the Web itself.[1] Web science research can therefore take numerous forms, seeking for example to gain new insights for political decision making and business practices. This could involve analyzing social media data on community and interactions; investigating user journeys, roles, and experiences with online services; and establishing guidelines around privacy, net neutrality, and security. For example, MIT's Internet Policy Research Initiative (https://internetpolicy.mit.edu) emphasizes public policy research, highlighting the way in which the Web science approach differs to that of data science.

Data science can be based on anything from the social sciences to healthcare and focuses on problems that arise from large amounts of data and the computational overhead required to process that data. Problems could require an investigation into what trends are present, or for a statistical model to be produced that can make predictions about future events; as a result, they often heavily employ quantitative methods. Many of these techniques are useful when carrying out a Web science study, given the scale of data available in that domain, and it is clear that the data science approach can be invaluable when assessing and planning Web development. However, data science's agenda goes much further; for example, Chris Mattmann discusses these techniques' use and importance in astronomy and Earth sciences,[4] areas in which the amount of data collected can reach phenomenal levels, and new techniques are required in order to make progress.

Data science clearly offers much to other disciplines' research agendas. Although Web science seeks to improve understanding about all of the Web's technical, social, and policy-based layers, data science offers new opportunities to gain insights into existing research initiatives in any discipline in which data is becoming—or has already become—prevalent. It seeks to develop the approaches to analyze this data, ensuring that it can be used to gain as much value as possible. The resulting techniques typically can be applied to numerous scenarios, increasing the potential for further value to be exploited elsewhere. Data science contributes new methods for knowledge discovery from ever-growing amounts of data[2]; the discipline's overall agenda therefore relates to supporting other disciplines to ensure that their datasets can be analyzed, exploited, and used to ensure that knowledge can be extracted, regardless of the data's size or complexity. As Web science does in relation to Web-related issues, data science seeks to ensure that policy decisions and business insights can be made across domains and that important scientific questions can be asked—and answered—from the growing stores of data that organizations of all sizes now have access to.

Given all this, data science clearly supports the ongoing development of Web science as a disciplinary lens through which to study the Web, and it helps to inform the design of new methodologies in the area—as network science does for techniques to study the underlying Web infrastructure and to understand social connections online. In order to answer the scientific questions posed about the Web, data science techniques are required to study the huge amounts of data available and extract the relevant knowledge. Owing to the rapid speed at which the Web is developing and changing, however, these techniques cannot stand still; therefore, the development of the data science agenda is driven by the changing needs of Web science, and other disciplines, which in turn are afforded new opportunities to ask deeper questions and gain more valuable insights.

Data Analysis

Data science tends to follow a general pipeline of data collection, processing, cleaning, analysis or modelling, and then visualizing and reporting. Within this, analysis typically occurs at two points. First, exploratory data analysis is carried out on the cleaned data to ensure that it is fit for purpose; statistics and observations are used to ensure that there are no duplicate entries or absurd outliers, for example, and calculations are made to describe the data's shape to ensure it meets expectations. Once this exploratory data analysis has been completed, the data scientist designs a model or algorithm to fit the data, depending on the type of problem and the question he or she is trying to answer. For example, this could be to classify a training dataset and develop a predictive algorithm to automatically classify future cases. Techniques include data mining to identify patterns and extract key relationships from the data variables and machine learning to improve predictive algorithms.

These techniques, understandably, also apply for large amounts of Web science research, particularly those employing the use of Web-based big datasets. Web science can employ data mining to notice trends in website usage, social media conversation, or information propagation. Longitudinal analysis of past behavior and machine learning can help model and predict future growth of a particular service or site or the future behavior of certain individuals. However, Web science can also go down more qualitative routes to carry out different types of analyses, employing coding and thematic content analysis to examine what is being said online, or using an ethnographic study to observe an online community's behavior. Web science is by no means limited to using just data science analytic methods, but they certainly provide an opportunity to use and exploit large amounts of data that could reveal trends and phenomena that would otherwise go unnoticed.

Observing the Web with Data Science

We have highlighted several ways in which Web science and data science differ, and we have shown where the two disciplines overlap and intersect—that is, the areas in which data science techniques and approaches produce the insights and results needed to complement other methods in a Web science context.

As part of the research resource being sourced by the Web Science Initiative, the Web Science Institute is developing a worldwide network of Web science observatories: a "global-distributed resource with datasets and analytic tools related to Web science."[5] The Web Science Trust website describes a Web observatory as "a system which gathers and links to data on the Web in order to answer questions about the Web, the users of the Web and the way that each affects the other" (see http://webscience.org/web-observatory/list-of-web-observatories). A Web observatory is, therefore, a global resource to provide both an archive of shared datasets and open analytics tools to facilitate Web science research.[8] As a simple definition, a Web observatory could be seen as a platform for carrying out data science on Web-based data, which often concern the Web itself.[9] Tools can be offered to support the entire data science pipeline of collecting, curating, managing, analyzing, and reporting on the data. Additionally, many studies about the Web are now becoming interested in real-time data analysis in which the data pipeline becomes a responsive and dynamic lens on Web data, allowing Web scientists to observe and respond to emergent social phenomenon. Addressing these needs, real-time Web observatories are being developed to offer a technical framework to archive and analyze data simultaneously, in an extensible manner.[10] For data science itself, several initiatives have aimed at producing similar resources; for example, IBM's ManyEyes, which ran from 2007–2015, allowed anyone to upload, visually analyze, and then share their data (www.computerworld.com/article/2930326/data-analytics/ibm-to-shutter-dataviz-pioneer-many-eyes.html). However, although Web observatories heavily emphasize reproducibility, data science offerings have focused on this less, so this is an area that can be exploited to improve the resources available to the discipline.

The Web observatory could offer an example of where data science can be used to gain insights into Web science research areas, as depicted by area 3 in Figure 1. Although the disciplines differ in scope, this intersection is an area in which the two could be used together to produce answers about the Web. As we mentioned earlier, a Web science study could focus on completely qualitative data, for which this approach would be less suitable, however for many studies—especially those containing quantitative social network data, weblogs, and longitudinal records of activity—the use of data science and an understanding of how to manage the volume and velocity of data involved could provide a Web scientist with the tools with which to study the Web. In other cases, data science approaches would not be suitable for the heterogeneity of the datasets often investigated in Web science—whereas in data science the data all tends to come from a particular subject domain, the interdisciplinary mix in Web science means that the data is often about a mix of concepts. Frequently, this involves data that is about people and their activities online, introducing a plethora of issues around privacy, security, and trust; therefore, rigorous access controls, ethical policies, and terms of conditions must be designed and implemented. These factors together introduce barriers to sharing data, which subsequently require a decentralized and distributed approach to solve—when using data from multiple organizations, it is often prohibited to move this data between servers. As an additional barrier, Web science data tends to be dynamic—not just from the perspective of the rate of creation, but also from the point of view of data management. Users can delete or remove data, which causes issues regarding the status of any collected or archived datasets; updates are required to keep them in line with these changes.

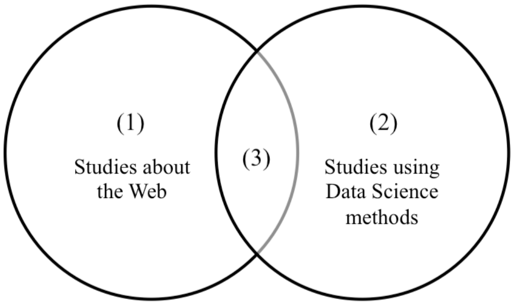

Figure 1. Overlap between Web science and data science research areas.

This discussion highlights some of the opportunities for data science within Web science observatories, but it also presents a range of important issues. These indicate areas where the two disciplines currently differ, but they also offer potential areas where the insights and best practices from Web science could be used to guide, influence, and inform the data science approach.

The Intersection Between Web Science and Data Science

Figure 2 displays the overlap of problem areas for which data science and Web science can be used together. The proceedings of the annual Web science conference are a broad mix of topics, including behavior analysis, social network analysis, social science, and studies from the humanities demonstrating that the scope of projects relating to the Web is itself massive. Data science could tackle further problems in which the data is not necessarily about the Web but uses Web technologies to store, manage, and distribute data in a way that ensures it can be used meaningfully.

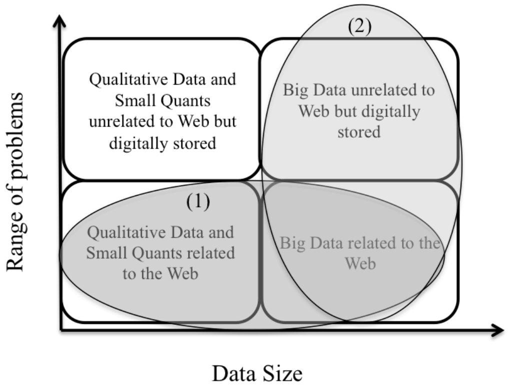
(2)
Qualitative Data and Small Quants unrelated to Web but digitally stored
Big Data unrelated to Web but digitally stored
(1)
Qualitative Data and Small Quants related to the Web
Big Data related to the Web
Range of problems
Data Size

Figure 2. Intersection of Web science (1) and data science (2) problem areas.

We have discussed the disciplines of Web science and data science as two distinct subject areas and have outlined the differences between them in terms of their approaches to and scope of study. However, there is a clear relationship that highlights areas in which these disciplines intersect and can complement each other significantly to achieve high-quality results in either field.

Data science can use Web science theory as a guiding influence on its approach to a particular problem. A key area in which data science goes beyond mere statistics is the domain knowledge that informs it and the actionable insights that are produced from the research design, which require an understanding of the context in which the results will be used. The use of a Web science approach to shape a data science study could reap rewards because the general strategy of mixing disciplinary perspectives to investigate the issue could provide the grounding and context to ensure that the data science itself both benefits the problem and produces results that can help alleviate the problem. For example, studying a large medical dataset about asthma symptoms requires expertise to be able to ask the correct question of the data, understand what the data is really showing, and ensure that the insights gained are accurate, validated, and verified by their existing domain knowledge.

A further opportunity for intersection and synthesis between these disciplines is for data science to be used as the quantitative element of a mixed-methods Web science study, in which the data in question is about the Web in some way and the data's volume or velocity makes it difficult to analyze without following the data science process. This approach might use resources such as a Web Observatory, which offers a platform for storing this type of data for use in research about the Web, and demonstrates the overlap between a science that seeks to establish new and evolving ways to handle data and one which seeks to study the very sociotechnical phenomenon responsible for producing much of the data itself.

References

1. S. Halford, C. Pope, and L. Carr, "A Manifesto for Web Science," *Proc. WebSci10: Extending the Frontiers of Society On-Line*, 2010; http://journal.webscience.org/297.
2. V. Dhar, "Data Science and Prediction," *Comm. ACM*, vol. 56, no. 12, 2013, pp. 64–73.
3. S. White and M. Vafopoulos, "Web Science: Expanding the Notion of Computer Science," *Proc. 43rd ACM Tech. Symp. Computer Science Education*, 2012, pp. 349–354.
4. C.A. Mattmann, "Computing: A Vision for Data Science," *Nature*, vol. 493, no. 7433, 2013, pp. 473–475.
5. T. Tiropanis et al., "Network Science, Web Science, and Internet Science," *Comm. ACM*, vol. 58, no. 8, 2015, pp. 76–82.
6. A. Wright, "Web Science Meets Network Science," *Comm. ACM*, vol. 54, no. 5, 2011, p. 23.
7. W. Hall and T. Tiropanis, "Web Evolution and Web Science," *Computer Networks*, vol. 56, no. 18, 2012, pp. 3859–3865.
8. T. Tiropanis et al., "The Web Science Observatory," *IEEE Intelligent Systems*, vol. 28, no. 2, 2013, pp. 100–104.
9. T. Tiropanis et al., "The Web Observatory: A Middle Layer for Broad Data," *Big Data*, vol. 2, no. 3, 2014, pp. 129–133.
10. R. Tinati et al., "Building a Real-Time Web Observatory," *IEEE Internet Computing*, vol. 19, no. 6, 2015, pp. 36–45.