The Search for the Data Scientist

Creating Value from Data

By Luis Felipe Luna-Reyes

The Internet of Things and Big Data trends are promoting changes in the professional and educational practices of experts in the fields of information, information technologies, computer science and operations management. Although new contents on big data and data analysis are being added to undergraduate and graduate courses, there is still the need of a better understanding of the skills, competencies and values of the Data Scientist. In this essay, I suggest a framework that includes five core sets of competencies to become an effective analyst and decision analysis steward. Skills of the data scientist are so diverse in nature that it is more likely that data science evolve from individual positions into a collective effort involving experts in different domain areas and levels of technical expertise coming from different organizational areas.

Keywords: Data Scientist, IoT, Smart City
Categories: *Social and professional topics~Computational thinking. Social and professional topics~Computing literacy.*
Corresponding Author: *Luis Felipe Luna-Reyes*
Email: *llluna-reyes@albany.edu*

Why a Data Scientist?

A couple of weeks ago, I attended a one-day workshop, and listened to current problems and concerns of Chief Information Officers of medium and small cities in the United States. Most of the conversation was around the increased challenges of managing and using data produced by personal devices and sensors across the cities. The context of the workshops was the new trend of the Internet of Things in the context of the Smart City movement, but as one of the participants commented, "the Internet of Things is about the data, not about the things." A recurrent topic during the session was the description of the human resources required to create value for the citizens by using the available data. What is the set of skills and competencies required by city managers and policy makers?

The problem is not new to information professionals. It was 1967 when Rusell Ackoff, in his classical paper *Management Misinformation Systems*,[1] pointed out that one of the main problems for managers and decision makers was not the lack of relevant information, but the excess of irrelevant information. Later, in 1982, John Naisbitt[2] coined the phrase "We are drowning on information, starved for knowledge," also pointing to the need of filtering

[1] R. Ackoff, "Management Missinformation Systems," *Management Science* 14, no. 4 (1967): 147–56.
[2] John Naisbitt, *Megatrends: Ten New Directions Transforming Our Lives* (New York: Warner Books, Inc., 1982).

and transforming existing information into actionable knowledge. Current trends of data warehousing, Internet of Things and Big Data had brought a renewed attention to the need of filtering and transformation data into relevant information or knowledge that can be used by decision and policy makers to solve organizational and social problems. As it has been pointed out in a recent reflection on Ackoff's original arguments, having massive volumes of data produced at high velocity and in a variety of formats has not only made the original problem worse, but also changed it in nature.[3] In this new environment, it is not only a matter of filtering from information generated through carefully designed information systems, but also a matter of finding new insights and patterns in unstructured data being produced in different formats and from many sources.

For Ackoff, the operations manager was the professional with the necessary skills to design the decision support systems that would use the appropriate models to better support managers in their decision making. The knowledge manager became the professional with the skill set to transform information into knowledge in the 90's and early 2000's. Since 2008, the profession of data scientist has emerged as the new professional in charge of helping managers, decision and policy makers to navigate the plethora of data to find insights, new applications and create value.[4] Although business programs have increased the number of data science related courses in an important way in the last years, there is still the need of clearly defining the skill set of a data scientist.[5]

What is a Data Scientist?

Given the urgency of processing, organizing and analyzing large volumes of data, current profiles of a data scientist fit two main categories, one with mastery of math, statistics and visualization techniques, and a second one with computer science and programing skills.[6] Some other more holistic approaches include in the profile of the data scientist some social science research methods skills –such as the ability to raise the appropriate questions and hypothesis--, as well as soft skills associated with communication and team work.[7]

---

[3] Kalle Lyytinen and Varun Grover, "Management Misinformation Systems: A Time to Revisit?," *Journal of the Association for Information Systems* 18, no. 3 (March 29, 2017), http://aisel.aisnet.org/jais/vol18/iss3/2.

[4] Thomas H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job Of the 21st Century," *Harvard Business Review* 90, no. 10 (October 2012): 70–76.

[5] Robert J. Mills, Katherine M. Chudoba, and David H. Olsen, "IS Programs Responding to Industry Demands for Data Scientists: A Comparison between 2011 - 2016," *Journal of Information Systems Education* 27, no. 2 (Spring 2016): 131–40.

[6] Michael Li, "What Kind of Data Scientist Do You Need?," *Harvard Business Review Digital Articles*, 2/1/2016 2016, 2–4.

[7] "What Is a Data Scientist? A Key Data Analytics Role and a Lucrative Career," *CIO (13284045)*, August 21, 2017, 5–5; Carlos Costa and Maribel Yasmina Santos, "The Data Scientist Profile and Its Representativeness in the European E-Competence Framework and the Skills Framework for the Information Age," *International Journal of Information Management* 37, no. 6 (December 2017): 726–34, https://doi.org/10.1016/j.ijinfomgt.2017.07.010.

Consistent to the more rounded perspectives of the data scientist, I want to advance a vision that includes five areas of skill relevant for data science, extending and better defining skill sets previously proposed in the literature:[8]

1. Computational Thinking: In the context of data analysis, computational thinking is related to the basic skills associated with the one role of the data scientist, developing applications and analyses that guide policy and decision making. It is possible to identify at least five modes of computational thinking:
    a. Computation as an algorithm.- As pointed out by some experts, a data scientist require knowledge of programing and application development in different development environments. Popular environments include, but are not limited to Java, C#, PHP, .NET, R or Python.
    b. Computation as decision modeling.- A model, in general, is a conceptual representation of a problem and it helps policymakers and other stakeholders structure the inquiry process. In many cases, data analysis requires the use of multiple analytical and modeling tools and techniques depending on the problem at hand.
    c. Computation as simulation.- One additional approach to computational thinking involves the use of simulation techniques. Some techniques involve system conceptualization as a set of interactive agents, and some other techniques focus on the system as a set of accumulations or states. Some simulation techniques are better suited for long term strategy, and others are more effective for optimizing operations or service delivery. Approaches of decision modeling and simulation start with a set of assumptions and hypotheses to be tested using data.
    d. Computation as machine learning.- One fourth mode of computational thinking involves data mining and machine learning techniques. Different from simulation and decision modeling approaches, machine learning techniques does not necessarily assume a theory or hypotheses as a starting point. On the contrary, machine learning techniques are most commonly used to identify insightful patterns or irregularities from the data without any initial assumption.
    e. Computation as statistical modeling.- Data scientists need also to be experts on the application of statistical tools to understand patterns and relationships

[8] Gabriel Puron-Cid, J. Ramon Gil-Garcia, and Luis F. Luna-Reyes, "Opportunities and Challenges of Policy Informatics: Tackling Complex Problems through the Combination of Open Data, Technology and Analytics," *International Journal of Public Administration in the Digital Age* 3, no. 2 (2016): 66–85; Jing Zhang, Luis Felipe Luna-Reyes, and Theresa A. Pardo, "Information, Policy, and Sustainability: The Role of Information Technology in the Age of Big Data and Open Government," in *Information, Models, and Sustainability: Policy Informatics in the Age of Big Data and Open Government*, ed. Jing Zhang et al., Public Administration and Information Technology (Heidelberg: Springer, 2016), 1–19; Costa and Santos, "The Data Scientist Profile and Its Representativeness in the European E-Competence Framework and the Skills Framework for the Information Age."

inside cross-sectional or time series data. The data scientist in this context requires understanding on the basic assumptions of each statistical method, as well as data requirements for each of them.

2. Domain Knowledge: Learning from data occurs in context-specific application domains. Each domain has specific requirements and problems that need to be clearly understood in order to choose the appropriate methods and processes for data analysis. Using data to solve transportation problems differs from using the same data to understand crime rates or emergency response. Specific domain knowledge is key to contextualize and reuse existing datasets at the city level.
3. Data Management: Managing and curating data, preparing it to use in current Internet of Things applications is still one key process. In other words, data exploitation and use to produce innovation and better policy and decision making requires careful data management processes in all the data life cycle from data creation to archiving or disposal. Data preparation and curation requires integration from multiple, disparate sources and usually need to be re-coded in a way that is consistent with the new application and context.
4. Enterprise Architecture: The data scientist requires knowledge about technology infrastructure and standards that facilitate the development of an enterprise architecture that enables data management and exploitation. Understanding the deployment of sensors, as well as the local and cloud-based services necessary to support the data management and exploitation process is a key skill to develop value from the data.
5. Stakeholder Involvement. Stakeholder involvement in the analysis process constitutes a key pillar to produce value for the citizens. Stakeholder engagement to facilitate data analytics modeling finds its roots in decision science, soft system approaches, and group decision support systems, and consists on a set of tools and techniques to elicit domain knowledge and values from stakeholders. Involving stakeholders in the process implies the skills and knowledge necessary to help them to pose relevant questions for their specific application domain, as well as other information requirements and hypotheses. Additionally, some techniques to engage stakeholders are used to build confidence on models, data analysis and policy development. Reporting and visualization skills are most important to build trust and commitment from stakeholders.

Conclusion

I would like to conclude this short paper with a conclusion and a brief reflection on the skills and competencies of the data scientist. An immediate reaction to the definition of core competencies suggests that it is unlikely that one single person will possess these diverse areas of expertise. As it has been pointed out previously in the literature, different organizations may need to emphasize on a skill-set when looking for a data scientist. The data manager or enterprise architect are maybe the most relevant ones when thinking about the organizational areas of information technology services of a city, and some form of

computational thinking and knowledge domain are the more relevant ones for a business area. Stakeholder involvement is a key skill set at higher levels when the position implies organizing working groups and collaborative teams.

In this sense, data science is a collaborative rather than an individual effort, and cities that are looking for the adoption and use of Internet of Things technologies to improve decision and policy making should consider not a single person to work as a data scientist, but to consider the creation of a team of professionals with the ability of working together to discuss and solve relevant problems to the city.

Additionally, and as it was discussed by CIOs in the workshop that I attended recently, the vision of a team is almost impossible to pull together in a small or medium city. However, potential models of collaboration that may allow cities to exploit available data involve the creation of city networks to share technical resources and applications. Finally, close collaboration with research centers and laboratories constitutes another way of acquiring the necessary skills and competencies to create value from sensor-produced data in smart cities.