Data science research in Sri Lanka: Human resource challenges and prospects[1]

Keynote presentation for South Eastern University, 10th Annual Science Research Sessions 2021, 30 November 2021

Rohan Samarajiva, LIRNEasia

Abstract

Based on the experience of building a data analytics unit within a Sri Lankan research organization since 2012, this presentation examines the challenges of doing cutting-edge data science in Sri Lankan conditions. Open-source software and relatively inexpensive hardware have lowered the barriers to participation. Constraints such as problems of access to data sets and funding are discussed briefly, with emphasis being placed on the challenges of recruiting and developing research personnel with the required skills. The field is subject to rapid change, with the even terms such as big data, now being replaced by newer terms. The importance of continued learning and working in interdisciplinary teams is highlighted.

[1] The comments of Sriganesh Lokanathan and Nisansa de Silva are gratefully acknowledged.

I was fortunate to have the opportunity to serve on the faculty of the Ohio State University, one of the largest state universities in the US, from 1987 to 2000. This was an exciting time. Search engines were beginning to transform the entire web experience. What had been once an esoteric insiders' club of academics and researchers was being made into the global public space it is today.[2] The socio-political and economic aspects of the Internet and the massive rise in data communication was beginning to attract the attention of researchers. I attended the first Computers, Freedom, and Privacy Conference, and owned a copy of the first Wired magazine.

We understood from the start that the study of the socio-political and economic aspects of the emerging technologies had to be inter disciplinary. As the recipient of a grant to get this conversation going, I found myself talking to senior professors at the Ohio Supercomputing Center, in the university's own Computing Center, in the Engineering Faculty, in the School of Public Policy and so on. One of the curious things I noticed was how many of these pioneers had PhDs in Chemistry. I asked around as to why.

The answer was staring me in the face. In the early days of computing, there were no computer science PhDs; there were no computer science BScs; there were no computer science departments. By necessity, everyone who was a pioneer in computing applications was a graduate in some other subject. At Ohio State, the scientists in the Department of Chemistry were writing computer programs for their research and had obtained the required hardware through the grants they obtained. They then rose in the various computer-related centers that they built up, and some of them gave up on chemistry. In other places, physicists may have taken on this role; in Sri Lanka it was electrical engineers; at Ohio State it was chemists.

These memories from the 1980s became relevant when I and Sriganesh Lokanathan, now Data Innovation and Policy Lead at Pulse Lab Jakarta (a joint venture of the UN and the government of Indonesia), ventured into research on big data applications and policy in 2012, almost 10 years ago. They are relevant today as well because researchers with the required skills and attitudes are the most significant constraint to data science research even at present.

Big data research

Back in the 1990s, I was working on what we now call "big data," but there were few at universities who could conduct research using these massive data sets. One needed access to supercomputers made by companies like Cray. These machines were so expensive and significant that the business pages reported individual sales; the number of countries that owned them could be counted on two hands. My research was on the policy implications of the novel capabilities made possible by these super-fast processors capable of handling massive data sets.[3] When I returned to Sri Lanka to work on telecom sector reforms in 1998, I closed off that line of research and focused on research relevant to local conditions.

In around 2009, reports of a new kind of data-based research appeared. An example was Google's efforts to track the emergence of seasonal flu outbreaks in the US by analyzing the terms used in billions

[2] http://openbookproject.net/courses/intro2ict/internet/history.html

[3] Burns, R.; Samarajiva, R.; Mukherjee, R. (1992). *Customer information: Privacy and competitive implications*, NRRI 92-11. Columbus OH: National Regulatory Research Institute; Samarajiva, R. (1997). Interactivity as though privacy mattered, in *Technology and privacy: The new landscape*, eds. P. E. Agre & M. Rotenberg, pp. 277-309. Cambridge MA: MIT Press.

of searches across the country. The claim was that the seasonal variations in search terms could tell decision makers which areas were experiencing flu outbreaks. Even though the findings were not based on a representative sample of the population, it was claimed that it was highly accurate.[4]

I wondered what kind of computing power they were using for that kind of near real-time analysis. But the pieces came together only in 2011 when I was listening to IBM Fellow C. Mohan speak in Colombo at an event organized by WSO2, a leading software company. I learned that supercomputers were no longer needed to analyze massive data sets, that major advances in storage memory allowed researchers significantly higher flexibility, and that the software was open source. We quickly put together a proposal to raise fund to conduct big data research of relevance to urban planning, a hot topic in the post-war conditions of 2011. I managed to negotiate access to pseudonymized mobile network data from multiple operators.[5]

The funders in Canada were very positive about the opening up of a new research front, but they had one question: did we have the people to do the research? It was a simple but decisive question. The funding decision rested on an adequate answer being provided.

Luckily, we had done some big-data-like research with a grant from the same funding agency a few years earlier. This was a collaboration with Auton Lab at Carnegie-Mellon University in the US[6] through which we sought to identify emerging diseases and propagation of infectious diseases.[7]

This work was different from what we describe as big data research in two aspects. One was that we were creating the data for analysis by positioning assistants next to doctors as they examined patients. This is very cumbersome and cannot be sustained over time, unlike in cases where the data is a by-product of whatever transactions that are occurring. The project gave good results but died as a result. An early report dubbed these data streams as Transaction Generated Information (more correctly, Transaction Generated Data).[8]

The second difference was that we were not doing the analysis ourselves. The lab at Carnegie Mellon had developed T-Cube, something they then described as “learning software” to analyze and predict faults in US Air Force aircraft. They wanted to see what other tricks could be taught to the software. Our data, with all the personally identifiable elements stripped out was run through T-Cube and would spit out analyses of various patterns. In today’s terms this was a machine learning application, that some would even describe as Artificial Intelligence. But it was embedded in that old paradigm: expensive supercomputers and proprietary software.

---

[4] Ginsberg, J., Mohebbi, M., Patel, R. et al. Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014 (2009). https://doi.org/10.1038/nature07634

[5] Some early results plus an account of how we went about gaining access to the datasets, how they were cleaned and analyzed, etc. may be found at Samarajiva, R.; Lokanathan, S.; Madhawa, K.; Kriendler, G., & Maldeniya, D. (2015). Big data to improve urban planning, *Economic and Political Weekly*, Vol L. No. 22, May 30: 42-48.

[6] https://www.autonlab.org/

[7] Waidyanatha,N.; Dubrawski, A.; Ganesan M.; Gow, G (2011). Affordable System for Rapid Detection and Mitigation of Emerging Diseases. *International Journal of E-Health and Medical Communications* 2(1). DOI:10.4018/jehmc.2011010105

[8] McManus, T.E. (1990*). Telephone transaction-generated information: Rights and restrictions*. Harvard University Program on Information Resources Policy. http://openbookproject.net/courses/intro2ict/internet/history.html

But the principal investigator on the project based on T-Cube was Nuwan Waidyanatha, a brilliant US trained mathematician with a Masters in Operations Research. He was no longer working fulltime for us, having moved to Kunming, China, but we could claim that he would be part of the team. The other key person was Sriganesh Lokanathan, who had a first degree in Computer Science from MIT and a master's in public policy from the Lee Kuan Yew School at the National University of Singapore. They would do the analytical work, while I would supply the policy expertise and serve as data wrangler. The funders accepted our response that we had a core team in place. The grant was approved.

Challenges

Once we obtained the funding, we found that Waidyanatha could not actually play a role, for various reasons, including location in China. The data were bound by strong non-disclosure agreements (NDAs) and could not be used outside our office. Lokanathan had the foundational knowledge in computing, but the languages and techniques used for analysis of big data in 2012 were not known when he was at MIT. This meant that he was scrambling to find people to conduct the research while rapidly educating himself on the required skills.

Continuous learning

We got lucky. Our first part-time researcher was Nisansa de Silva, a computer science graduate from the University of Moratuwa, who was both a hard worker and a fast learner. Through him, we were able to recruit a core team of University of Moratuwa graduates who wanted to do cutting-edge research and get the publications that would enable them to enter good post-graduate programs. Now they all have PhDs or all about to get them. Dr de Silva is back as a Lecturer at the University of Moratuwa. This first set had a solid foundation in computer science, but had to learn data analytics while working. We also established working relationships with Joshua Blumenstock, perhaps the leading researcher using big data to generate development-related insights, and with colleagues from the MIT economics program.

Beyond the initial team, we recruited many interesting researchers with varied backgrounds. Aparna Surendra is illustrative. When she came to us wanting to work as an intern, having no formal qualifications in computing; she was an English Major with work experience in international relations. But she was already taking online courses in data science and was from Stanford, where the very air people breathe seems to include computing. She did excellent work with us,[9] went on to get a placement as an intern at Deep Mind, one of the world's most prestigious artificial intelligence companies, and is now Senior Associate at AWO Agency, a leading data rights organization that blends data science, law and ethics.

Another star performer with an unusual career path was Dedunu Dananjaya, now a database engineer at WISE, based in Estonia. He was recruited based on a blog post he had written on data analytics. He was outside the formal credentialling system, mostly self-taught and in the process of completing an external Bachelor's in IT degree program offered by the University of Colombo. In his case, online learning was complemented by interactions with those who had been formally trained in programming. There was no doubt about his intelligence and talent, but the rounding out he received as a result of

[9] Fernando, L.; Surendra, A.; Lokanathan, S.; Gomez, T. (2018). Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka*. DSMM'18: Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*: 1–12. https://doi.org/10.1145/3220547.3220549

working with formally trained computer scientists helped make him one of the most valued members of the data science team.[10]

Big data was the focus of work back in 2012. But now the field has moved on. Everyone is working on what is colloquially known as artificial intelligence. That means that whatever our current researchers learned while at university is not enough. In the time sheets that all researchers complete so that their time can be billed to various projects, we introduced a column for time spent on learning. Everyone at LIRNEasia is expected to engage in learning; we not only allow paid time to be used for learning but in some cases, we will pay for the courses and the credentials. We found that the researchers in the team that started off as big data (now called Data Algorithms and Policy) spent 25 percent of their time learning, double what other researchers did.

What this means is that the content of what was learned in degree programs prior to joining us matters less than the willingness and the ability to learn. Some knowledge of computing does help, as we have seen from some of our failure cases. But even the self-taught, like Surendra and Dhananjaya, can achieve excellent results. It appears that things are not much different from what was happening at Ohio State back in the 1970s, when people from various disciplines were boot-strapping themselves into computer scientists.

Interdisciplinary teams

Most, if not all, data science papers have multiple authors.[11] What we found was that the data science people had to know how to speak the language of the domain experts in whatever field they were working in. For example, our own work on dengue propagation,[12] included medical researchers; a paper on transportation included a transportation specialist.[13]

This is never easy, but possibly the US and Canadian university practice of requiring undergraduates to take courses outside their specialization provides an advantage to those who come from such systems as against the rigidly disciplinary curricula found in Sri Lanka. But it may be possible for newer programs to build in more opportunities for inter-disciplinary learning. What is required is the ability to communicate across disciplinary boundaries and to understand what those with domain expertise are saying, rather than expertise in multiple domains.

---

[10] Lokanathan, S.; Kreindler, G.E.; de Silva, N.H.N.; Miyauchi, Y.; Dhananjaya, D.; Samarajiva, R. (2016). The potential of mobile network big data as a tool in Colombo's transportation and urban planning, *Information Technology and International Development*, 12(2): 63-73. http://itidjournal.org/index.php/itid/article/view/1506

[11] For an example of non-LIRNEasia originated influential data science research see: Wesolowski, A.; Qureshi, T.: Bonid, M.F.; Sundsøy, P.R.; Johansson, M.A.; Rasheed,S.B.; Engø-Monsen, K.; Buckee, C.O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan, *PNAS*, 112 (38) 11887-11892. https://doi.org/10.1073/pnas.1504964112

[12] Dharmawardana, K. G. S., Lokuge, J. N., Dassanayake, P. S. B., Sirisena, M. L., Fernando, M. L., Perera, A. S., & Lokanathan, S. (2017). Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data. *Proceedings of the 12th IEEE International Conference on Industrial and Information Systems* (ICIIS 2017). https://lirneasia.net/2017/12/predictive-model-dengue-incidences-sri-lanka-mobile-network-big-data/

[13] Maldeniya, D., Lokanathan, S., & Kumarage, A. (2015). Origin-Destination matrix estimation for Sri Lanka using mobile network big data. 13th International Conference on Social Implications of Computers in Developing Countries. Negombo. https://lirneasia.net/2015/05/origin-destination-matrix-estimation-sri-lanka-mobile-network-big-data/

Our most successful data scientists have been those who were able to think beyond disciplinary boxes and read widely. I recall mentioning a book club organized outside office hours in my recommendation letters for one of our researchers who is completing a PhD in computational social science in the US.

Recruitment never stops

It is well known that data scientists are in short supply and that they are paid well. Even when the salaries are high, it is difficult to hold on to them. The strategy at LIRNEasia was to think of the compensation in terms of a package that included the opportunity to do innovative work and publish leading to higher probability of gaining funded admission to high-profile PhD programs. It is thus normal for researchers to cycle through.

This requires continuous recruitment and mentoring. For effective recruitment and mentoring it is necessary for there to be a stable core leadership team. When that condition cannot be maintained, the entire operation is weakened. Especially in the case of team members recruited with unconventional backgrounds, it is critically important to have in place strong mentoring and advising capabilities in the core team. Otherwise, it is possible for poor quality research practices to become entrenched in the organization.

Funding

Research requires adequate levels of funding. Sri Lankan universities are limited in what they can pay researchers who work on funded projects. In addition, funding from outside the country now requires Cabinet approval, which is said to take inordinate time, given the layers of prior approvals required. These problems are likely to seriously disadvantage university-based data science research.

Access to large datasets

One must have big data to do big data research. This is especially important in the context of the heightened importance of machine learning. Obtaining access to datasets that are under the control of government agencies is difficult because of attitudes which are hostile to sharing and also because the quality of the datasets is problematic. For example, see the transport data on the government's open data portal,[14] where the most recent data is nine years old.

LIRNEasia did manage to gain access to valuable pseudonymized datasets from the private sector. The difficulties experienced are described in detail elsewhere.[15] With apparent adoption of the rather rigid European model of data protection,[16] it may become even more difficult for outside researchers to gain access.

If all data use has to be covered by consent and purpose limitation principles, it may not be possible to permit use by third parties for traffic management, energy management, urban planning etc., since these uses could not be conceptualized at the time of signing up customers. So, what is likely to be the result of mechanical extension of inform and consent rules that were developed for qualitatively different conditions of the past would be giving the big companies such as mobile operators or search

[14] http://data.gov.lk/search/field_topic/transport-22

[15] Lokanathan, S.; Madhawa, K.; Kriendler, G.; Maldeniya, D. (2015). Big data to improve urban planning, *Economic and Political Weekly*, Vol L. No. 22, May 30: 42-48.

[16] Personal data protection, a bill, *Gazette of the Democratic Socialist Republic of Sri Lanka*, Part II of November 19, 2021.

providers a monopoly on large datasets of transactions; and shutting out small firms and public interest users, including university researchers.[17] However, there may be other datasets such as satellite data that are either public, or available for purchase, that can be used for research.[18]

Prospects

The heyday of data science is actually behind us. For most purposes, what used to be done under the rubric of data science is now being done as AI research. In addition, completely new strands of research are being opened up such as Web 3.0.[19] It is unlikely that there will ever be a situation where what is taught in universities can be directly applied in cutting-edge research in any of these areas.

Universities will have to establish good working relations with research organizations and firms so that their students can learn what is happening at the cutting edge, while equipping them with foundational knowledge which is what universities are good at. In that sense, the story told above of a particular experience of setting up a big data research unit in Sri Lanka has broad relevance.

[17] https://lirneasia.net/2013/10/big-data-for-big-boys/

[18] Shanmugarajah, Y.; Chandana, M. (2020). The State of the Art in Leveraging Public Domain Remote Sensing Data for Development Purposes. https://lirneasia.net/2020/05/the-state-of-the-art-in-leveraging-public-domain-remote-sensing-data-for-development-purposes/

[19] Mak, A. (2021 November 9). What Is Web3 and Why Are All the Crypto People Suddenly Talking About It? *Slate* https://slate.com/technology/2021/11/web3-explained-crypto-nfts-bored-apes.html