The Future of Academic Statistical Consulting Centers in the Era of Big Data

Osama A. Hussien
Department of statistics
Alexandria University Egypt

Mathematics Subject Classification 62A01 · 62H99

Key Words:

Statistical Consulting; Big Data; Data Sciences; Data Reproducibility; result Replicability; Statistical Methodology; Machine Learning.

ABSTRACT

In the age of big data and data sciences statistics made a real difference and reinforced the impact of statistical thinking on society. The demand for statisticians has steadily grown worldwide. . The pace of supply and demand for statisticians in the Arab states has not yet matched international speed, partly due to the lack of a clear appreciation for the role that statistics plays in decision-making and fostering critical thinkers. As a result a widespread misuse and inefficient use of statistics in scientific research plus anti-statistics feeling and distrust of findings which rely on statistical evidence. Moreover, data reproducibility, analysis reproducibility/stability, and result replicability are critical for improving accuracy and transparency in scientific research, especially when dealing with big data. Part of the solution to these problems is the establishing of academic statistical consulting centers. Strong academic statistical consulting centers could be a bridge to the scientific community of scholars and all users of data sciences methods to convey strong value-added practice of statistics, and the infusion of statistics into interdisciplinary research. This paper discusses a proposed mission of academic statistical consulting centers, in Arabic states, to coup

with the new challenges (and opportunities) of the new fast growing paradigm of data sciences.

1-INTRODUCTION

Statistical consulting is a professional collaboration of a statistician with another professional to improve research by improving the quality of the statistical aspects of research, Gibbons, J. and Freund, R. (1980). Statistical consultation could be more of a collaborative/partnership character, where the statistician is a member of a team. The statistician then invests a lot of time and effort, to become knowledgeable in the subject matter area and expert in the applications of statistical methods in that area, Sundberg (2011). The academic statistical consulting units are not likely to be viewed as legitimate academic units if the university community believes its role should be primarily one of service to non-statistics faculty. So it is important to define and justify ideal roles and objectives of proposed academic consulting units, especially in Arab universities, Carter et al. (1986). Providing a good statistical consultancy service depends on many factors. The first is general factors influencing the training of consultants, the status of consultancy and the organization of the service. The second is associated with the role of the consultant in an experimental team and his relationship with other members of the team, Sprent (1970). The third related to the "scientific culture" environment surrounding scientific research. In developing countries, the third factor is a critical obstacle to the development of scientific research; see Badr (2018a and 2018b), Rached and Craissati (2000) and Perrino et al. (2013). Our science doesn't exist in a vacuum. Statistical aesthetics depend upon culture. More importantly, the rapid growth of big data and aspects of data sciences leads to new methods of statistical analysis and new areas of research and academic statistics consultant should be prepared for the new era. The science of statistics needs to build new theory and methods to meet the current and future challenges of data science. At the same time keeps the end-goal of statistical analysis of drawing conclusions based on the quantification of uncertainty. This paper discusses a proposed mission of academic statistical consulting centers, in Arabic states, to coup with the new challenges (and opportunities) of the new fast growing paradigm of data sciences.

Thus the aim of this paper is to discusses a proposed mission of academic statistical consulting centers, in Arabic States, to coup with the new challenges (and opportunities) of the new fast growing paradigm of data sciences.

2. CURRENT STATE OF ACADEMIC STATISTICAL CONSULTING CENTERS

In Arabic states the universities are structurally divided into several colleges, each with some needs for statistical consulting and cooperative research services. Typically the department of mathematics (college of science) consternate on theoretical statistics and probability theory while the department of applied statistics (college of business) concern with statistical business and economics applications). Moreover scattered statistical users exist at other colleges, e. g. agriculture, education and public health. It would seem logical and more efficient to establish one strong statistical consulting center of institutor status with a director reporting to a university administrator at the vice-president level, preferably the Vice-President for Research. Such an administrative structural arrangement should provide a maximum opportunity for the center to serve all the statistical consulting needs in an impartial manner, Bancroft (1974). The existing university statistical consulting centers in Arabic universities are not interdisciplinary collaborators. In most of the cases it is a branch of the department of statistics to provide advice and/or assistant in statistical analysis for faculty and graduate students. In some cases it is a branch of business consulting centers. The Statistical Consulting Unit at Qatar University is a good example of a collaborative university statistical consulting center. They offer individual collaboration meetings, short courses, and support for interdisciplinary research projects. Another example is the applied statistics center at Beirut Arab University. They have a clear mission of assisting students, faculties, and private sectors in the "data analysis process". The Institute (Faculty) of Statistical Studies and Research, Cairo University established its statistical consulting center in 1994. They have a good vision of enriching statistical culture, enlightening society, and unveiling a kind of statistical illiteracy. The center used to provide distinguished statistical consultations in different scientific fields. But the average number of consultations provided in a year is 200 consultations, while the number of Ph.D. and M.Sc. degrees offered at Cairo University average 5100 a year. (see http://srv2.eulc.edu.eg/eulc_v5/libraries/Thesis). This means a high percentage of the statistical consultations for are done in private centers. Center for Surveys and Statistical Applications has been established in 2002, as a statistical research unit embodied in the Faculty of Economics & Political Science; Cairo University. The mission of the center is to raise awareness of the staff in governmental and non-governmental organizations about

population, health and education problems and to empower the research community in conducting the statistical studies and applications; especially those related to the society issues. The main activity of the center is conducting surveys about the current demographic issues such as health, woman, and education. Other statistical consulting centers were established in King Saud, King Abdulaziz, King Fisal Om Elkora, Kuwait, Emiratis and other Arabic universities. All such centers are not interdisciplinary collaboration consulting center. The impact of this situation is a decline in research as a whole and lowering the academic rank of the university.

The "scientific culture" environment surrounding scientific research in developing countries can be summaries in the following:

- **Lack of research skills in modern methods**
- **Lack of equipment for carrying out state- of- the art research**
- **Inadequate research funding**
- **Overloaded teaching and administration schedules which leave little or no time for research. Awe (2020).**
- **Lack of collaborative research, data sharing, data synthesis and scientific equity, Perrino1 et al (2013).**
- **The nonprofessional statistical analysis done by private consulting centers who have a big share of the statistical consulting centers in the area.**
- **Lacking the dual purpose of training statistics students and providing statistical support to researchers.**
- **The wide spread of data sciences applications done by computer scientist ignoring the scientific statistical methodology.**

3- SCIENTIFIC RESEARCH AND THE CHALANGE OF BIG DATA

The statistical challenges presented by the widespread use of data science tools are growing increasingly. Much of data science has focused on purely predictive ''black box'' tools rather than classical modeling, inference, and analysis. Observations are often made without proper experimental design, resulting in biased and incomplete data. The populations studied have a high degree of heterogeneity Modeling and inference procedures specifically designed for these types of scenarios are desperately needed if data science is ever to be put on a firm inferential footing. Big data are often associated to the idea of *data-driven* research, where learning happens through the accumulation of data and the application of methods to extract meaningful patterns

from those data. This new shift in paradigm to data-driven research re-focuses the emphasis away from raw computational power to the development of specialized algorithms for learning from data and reasoning with the data. Thus Machine Learning (ML) techniques have come to play a central role in automatic data processing and analytics across a wide spectrum of research domains. However, the lack of well-defined principles in choosing ML algorithms suitable for a given problem remains a major challenge. See Schaeffer (2002), Berman and Crosas (2020) and Berthold (2019).

Within data-driven inquiry, researchers are expected to use data as their starting point for inductive inference, without relying on theoretical preconceptions, in contrast to theory-driven approaches where research consists of testing a hypothesis (Anderson 2008, Hey et al. 2009). As a discipline that deals with many aspects of data, statistics is a crucial milestone in the rapidly evolving data sciences. Statistics should play a leading role in data science by assisting with the use of data and decision making in the face of uncertainty without sacrificing the scientific statistical methodology .

Data science usually thought to be 80% data preparation and 20% analysis. Good data scientists should be 20% data preparation and 80% analysis. The automated methods for complex data analysis lack of consideration of (1) interpretability, (2) uncertainty quantification, (3) applications with limited training data, and (4) selection bias. Statistical methods can achieve (1)-(4) with a change in focus, (Olhede and Wolfe 2018, Galeano and Peña 2019, Meng 2018 and, Reid 2018 . These issues should be seriously considered in Big Data analysis and in the development of statistical procedures.

As any scientific research report, a data science application report should be all of these things:

- **Meaningful: (not necessarily conclusions about the population).**
- **Measurable: metrics should be clearly defined and agreed upon by the analytical experts on the team.**
- **The three R's of Science: Reproducibility; Repeatability and Replicability (Reliability). These concepts are pivotal for improving rigor and transparency in scientific research, especially when dealing with big data (National Academies of Sciences, Engineering, and Medicine, 2019). This includes data reproducibility, analysis reproducibility/stability, and result**

replicability. Recently, much attention has been focused on the replicability of scientific results, causing scientists, statisticians, and journal editors to examine closely their methodologies and publishing criteria. Reproducibility refers to the ability to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator (Meng 2020). Replicability, on the other hand, refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected (Daoudi et al. 2021), and generalizability refers to the extension of the scientific findings to other populations. Benjamini (2020) and Marinez et al. (2021).

- **Generalizability: refers to the extension of the scientific findings to other populations, contexts, and time frames.**

A successful statistics and data science applied research requires a wide range of skills, including domain knowledge, analytical thinking , problem-solving, teamwork , project management, communication, machine learning, big data, and software development.

4- Extending the mission of academic statistical consulting centers as an Engine for Research developments

The new fast growing paradigm of data sciences have led to a paradigm shift in the conceptualization of statistics teaching and training. This involves reforms of the graduate and undergraduate curriculums, training and qualification of the statistical scholars, in both teaching and research, more real data projects for student's graduation and establishing interdisciplinary collaboration in teaching and research with other disciplines. Accomplish those goals needs hard effort, financial funding and long time. This will be hard in most of the developing countries. This is partly due to the lack of a clear appreciation for the role that statistics plays in decision-making and fostering critical thinkers (Innabi 2014). Another important factor is the misunderstanding of data science as a branch of computer sciences that produces algorithm to solve almost all applied problems.

Strong academic statistical consulting centers (ASCC) could be a bridge to the scientific community of scholars and all users of data sciences methods to convey strong value-added practice of statistics, and the infusion of statistics into interdisciplinary

research. The ASCC have fewer restrictions than the corresponding academic departments. They can teach new subjects, may have suitable financial funding and its mission is essentially a collaboration with "all" scientific communities in the society. The ASCC have the chance achieve (with the corporation of other identities through practical research projects) analysis reproducibility; stability, and result replicability Improving the mission of the ASCC, they can provide efficient and actionable solutions to complex big data problems. The missions we propose are extensions and modifications to LISA2020 missions. ASCC should become an excellent research center that lead the scientific research and applied projects. It will be a bridge between business and industry and the academic community. Highly Principled Data Science insists on methodologies that are: (1) scientifically justified; (2) statistically principled; and (3) computationally efficient. The following missions for ASCC are proposed.

Mission 1: Bridge the concepts in data sciences and statistical sciences by improve statistical and data science skills.

- **Become a center of expertise and excellence in quantitative and qualitative data analysis methods and relevant computer packages.**
- **Produce a number of research reference guides covering all aspects of the research process.**
- **Teach short courses and collaborative joint workshops to bridge the concepts in data sciences and statistical sciences. This may include subjects like: data structures and algorithms ; scientific computing ; optimization techniques; data visualization; machine learning; deep learning; exploratory data analysis; experimentation and evaluation tools; modeling in data science and statistics and scientific and statistical significance methodologies.**

Mission 2: When Models Meet Data.

Training the collaborative statisticians and data scientists to balance the needs of theory, algorithms and practice to get reproducible; stabile, and replicable results for real-world problems.

- **Providing assistance with project planning and writing of research proposals.**

- **Providing a statistical advisory service to facilitate data acquisition, capture and analysis within research projects.**
- **Promote the use of Information Technology in scientific applications and projects to assist in the creation of significant information networks that can foster intellectual collaboration.**
- **Conceive researcher (equipped with big data) that research idea are based on the research hypotheses not on the availability and the size of his data.**
- **Expert statistical consultant can have a significant impact on the research design and decisions based on their ability to influence their peers.**

- **Promote the use of Information Technology in scientific applications and projects to assist in the creation of significant information networks that can foster intellectual collaboration.**
- **Conceive researcher (equipped with big data) that research idea are based on the research hypotheses not on the availability and the size of his data.**
- **Expert statistical consultant can have a significant impact on the research design and decisions based on their ability to influence their peers.**

Mission 3: Efficient Research infrastructure.

Create a space for collaboration to transfer academic evidence into action.

- **The big data era has created a new scientific paradigm: collect data first, ask questions later. The scientific (statistical methodology) set the hypotheses first, and sometimes propose a model, before collecting "random" data sample.**
- **To bridge the conceptual and methodological gap it is essential that both communities should work together, not as consultants and clients, but as genuine partners and co-investigators in scientific investigations.**
- **To make this partnership truly effective, and mutually beneficial, will require investing time and energy on both sides to understand each other's language, and perspectives, and modus operandi.**
- **Statistical consultation could be more of a collaborative/partnership character, where the statistician is a member of a team, and the aims are more far-reaching.**

- **Statistical leadership in a collaborative environment is the use of influence without authority to guide the design, strategy, and decisions of a multidisciplinary team.**
- **Multidisciplinary research teams are a network of relationships that are not hierarchical in nature. But, statistical leadership can have a significant impact on the research design, the analysis procedures and results if they can build a healthy environment of collaboration.**

5-NECESSARY SKILLS FOR STATISTICS AND DATA SCIENCE CONSULTANTS

The ASCC needs to train its members to new skills. This should include interpersonal skills; self-management skills, collaboration skills and networking, cultural understanding, and technical (methodological and computational) skills. The technical skills should include recent advances in data visualization, machine learning, data mining and high-performance computing along with the critical statistical concepts for scientific research like randomness, clean data, significance models and scientific inference. At the same time the statistical consultant need to compromise between a perfect analysis and practical methods. In big data type problems the "new" statistical consultant will teach (and convince through corporation) the client critical statistical concepts for scientific research like randomness, clean data, significance models and scientific inference.

Consulting skills

1-Gaining Interest in Consulting

2. Strengthening Technical Skills

3. Understanding and Interacting with Clients

Collaboration skills

1-Communicating with impact to a multidisciplinary audience is a key to successful consultation.

2-Recognizing the Importance of Collaboration to Research Practice

3-Developing Relationship Skills

4-Organize your thoughts and ideas before meeting with the research team.

5- Stimulate the team members to speak up and share ideas.

6- Give members of the research team a chance to hear their input in a manner they can understand.

Leadership skills

Under statistical leadership the multidisciplinary collaborative team will be able to solve many technical challenges and deliver practical scientific inference and predictions that can be reproducible and stabile. The keys to statistical leadership are competency in listening, networking, and communication, Gibson (2019).

- **Active listening is a process that requires the statistician to explain their perception of the problem and ask various members of the research team to identify any misconceptions or additional relevant information to ensure an accurate understanding of the problem from multiple viewpoints.**
- **The multidisciplinary team is a network of relationships which need to be proactively developed to facilitate successful statistical leadership.**
- **The consulting training needs to cover the whole data developing process.**

Technical skills: (methodological and computational)

Statistical consultant should incorporate advances in data visualization, machine learning, data mining and high-performance computing along with the critical statistical concepts for scientific research like randomness, clean data, significance models and scientific inference. The consulting training needs to cover the whole data developing process. At the same time the statistical consultant need to compromise between a perfect analysis and practical methods. In big data type problems the "new" statistical consultant will teach (and convince through corporation) the client critical statistical concepts for scientific research like randomness, clean data, significance models and scientific inference.

An example of the new technical skills are the methods to "Data integration by combining big data and survey sample data for finite population inference" presented by Yang and Kim (2020) and Kim and Tam (2020). Another example is sampling methods from big data with uncertainty by He et al. (2014).

REFERENCES

1. **Anderson, C.(2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, 23 June 2008.**
2. **Awe, O. (2020). Virginia Tech's LISA 2020 Program: Evidence from Nigeria. www.stat.lisa.vt.edu/LISA2020.**

3. **Awe, O., Crandell , I. and Vance, E. (2015). Building Statistics Capacity in Nigeria Through the LISA 2020 Program. ISA.**
4. **Badr, M. Z. (2018). Challenges Facing Scientific Research in Developing Countries: 1. The Human Factor. Egyptian Journal of Basic and Clinical Pharmacology, vol. 8, Article ID 101378, 3 pages, 2018**
5. **Badr, M. Z. (2018). Challenges Facing Scientific Research in Developing Countries: 2. Environment and Resources. Egyptian Journal of Basic and Clinical Pharmacology, vol. 8, Article ID 101388, 2 pages, 2018.**
6. **Bancroft, T. (1974) On Establishing a University-Wide Statistical Consulting and Cooperative Research Service. The American Statistician, 28, 21-24.**
7. **Benjamini, V. (2020). Selective Inference: The silent killer or replicability. Harvard Data Sciences Review,2-2.**
8. **Berman, F. and Crosas, M. (2020). The research data alliance. Benefits and challenges of building a community organization. Harvard Data Sciences Review, 2-1.**
9. **Berthold, M. (2019). What does it takes to be a successful data science. Harvard Data Sciences Review, 1-4.**
10. **Carter, R. , Scheaffer, R. and Marks, R. (1986) The Role of Consulting Units in Statistics Departments The American Statistician,40, 260-264.**
11. **Daoudi, N., Allix. K., Bissyand, T. and Klein, J. (22021). Lessons Learnt on Reproducibility in Machine Learning Based Android Malware Detection. Empirical Software Engineering (2021) 26: 74.**
12. **Gibbons, J. And Freund, R. (1980) Organizations for Statistical Consulting at Colleges and Universities The American Statistician, 34, 140-145.**
13. **Galeano, P., Peña, D. (2019). Data science, big data and statistics. TEST 28, 289–329.**
14- **He, Q. , Wang, H. , Zhaung, F., Shang, T. and Shi, Z. (2014). Parallel sampling from big data with uncertainty distribution. Fuzzy sets and systems, 1-14.**
15- **Hey, T., Tansley S, and Tolle K.(2009). The Fourth Paradigm. Data-Intensive Scientific Discovery, Redmond, WA: Microsoft Research.**
16- **Innabi, H. (2014), Teaching Statistics in the Arab Countries: The Ambitions and the Needs. Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9), Flagstaff, AZ.**
17-**Olhede, S. and Wolfe, J. (2018) ,The future of statistics and data science. Statistics and**

Probability Letters 136 , 46–50.

18-Perrino, T., Howe, G., Sperling. A., Beardslee, W., Sandler, I., Shern, D., Pantin, H., Kaupert, S., Cano, N., Cruden, G., Bandieraand, F. and Brown, H. (2013). Advancing Science Through Collaborative Data Sharing and Synthesis, Psychological Science, Vol. 8, No. 4 433-444.

19- Rached E. and Craissati D. (2000). Research for Development in the Middle East and North Africa. international development research center. Ottawa.

20- Reid, N. (2018). Statistical science in the world of big data. Statistics and Probability Letters 136, 42–45.

21- Scheaffer , R. (2002) , Statistical Bridges Journal of the American Statistical Association , 97, 457- 1-7.

22-Sprent, P. (1970). Some Problems of Statistical Consultancy. J. R. Statist. Soc. A,133, 139-165.

23- Sundberg, R. (2011). Statistical Consulting .International Encyclopedia of Statistical Science,1390-1391.

24- Kafadar, Karen (2020) Reinforcing the Impact of Statistics on Society. J. Amer. Statist. Assoc., 115, 491–500.

25-Kim, J, and Tam S. (2020) Data integration by combining big data and survey sample data for finite population inference. International Statistical Review 89(2) 382–401.

26-National Academies of Sciences, Engineering, and Medicine (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press.

27-Locascio, J. (2019) The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration The American Statistician, 73, 346–351.

28-Meng, Xiao-Li (2020). Reproducibility, Replicability, and Reliability. Harvard Data Science Review. 2-4.

29-Meng, Xiao-Li (2018). Conducting highly principled data science: A statistician's job and joy. Statistics and Probability Letters 136, 51–57.

30-Martinez, I. Viles, E. and Olaizola, I. (2021). Data science methodologies: current challenges and future approaches. Big Data Research, 24,100183.

31-Tatiana Perrino, George Howe, Anne Sperling, William Beardslee, Irwin Sandler,David

Shern, Hilda Pantin, Sheila Kaupert, Nicole Cano, Gracelyn Cruden, Frank Bandieraand C. Hendricks Brown (2013) Advancing Science Through Collaborative Data Sharing and Synthesis, Psychological Science, Vol. 8, No. 4 433-444.

32-Vance, E. (2015). Recent Developments and Their Implications for the Future of Academic Statistical Consulting Centers. The American Statistician,67,127-138.

33-Yang S. and Kim J. (2020). Statistical Data Integration in Survey Sampling: A review. Japanese Journal of Statistics and Data Science, 3, 625—650.