Data Science: Profession and Education

Longbing Cao
Data Science Lab, University of Technology Sydney

***Keywords*—Data science, advanced analytics, big data analytics, data science profession, data science courses, data science education**

Advanced analytics, data science, and new-generation artificial intelligence (AI) are among the most promising directions in the area of information and communications technology (ICT) and the disciplines of science, engineering and technology (SET), where data science has become the driving force of new-generation AI. Data science and new-generation AI have attracted increasing interest from governments, companies and academia, with important initiatives launched by countries such as the United States [1] and China [2], and the European Commission [3].

However, despite the fact that the role of a data scientist has been described as the sexiest job in the 21st century [4], [5], the qualifications and capabilities of a data scientist are not clearly defined; it is important to determine and define the qualifications needed by next-generation data scientists who will be responsible for transforming science, technology and innovation both today and in the future and to increase economic competitiveness [6].

An increasing number of data science roles have been generated by increasing online course offers and face-to-face learning in traditional institutions [7]. Hence, data science has clearly become a highly sought-after profession [8], [9], [10]. However, high-end data science employers (including innovators, vendors, and enterprises) often complain about the limited availability of qualified data scientists to enable their strategic development and foster their future competitive advantage. This indicates the changing status of existing professional and educational markets, the restricted benchmarking and accreditation of the responsibilities and capabilities of data scientists, and the urgent need to standardize and upgrade competencies and the maturity of data science qualifications and education. This article addresses these important issues, with the aim of contributing to the standardization and formalization of the next-generation data science profession and education.

I. Data Science As A Profession

Data science has driven the emergence of a new profession: the *data science profession*, or simply the *data profession* [6]. Evidence for the growth of the data profession includes the increasing number of clearly divided, diversified, well-defined and predominant data-oriented roles and responsibilities, the increasingly clearly articulated and pursued qualifications, the increasingly clarified competencies and capabilities of these roles, and their increasingly profound impact on driving and transforming data research, innovation, the economy and society.

A. Spectrum of Data Roles

Data roles broadly refer to job positions and the responsibilities that are centered on, related to, or enabled by data and data-oriented (including data-driven, data-enabled, and data-based) tasks and agendas. Specifically, *data science roles* are those centered on data-driven discovery, innovation and practices. They have been built on the profession of software engineering, database administration, business intelligence, computing, enterprise application integration, business analysis, and statistical analysis.

There are various perspectives from which to structure and categorize data roles. To conduct enterprise data science, an enterprise data science team is often formed with different roles to accomplish various scientific and engineering tasks. Typically, the members of a data science team play different data roles to undertake the following responsibilities and tasks: (1) to design the *infrastructure* for data computing, analytics, and decision-support; (2) to perform the *engineering* of data, software, applications, networking, communication, analytics, reporting, and decision-support; (3) to undertake the *discovery* of data-driven scientific exploration, modeling and optimization, including general and specific analytics, mining, learning, recognition, prediction and refinement; (4) to develop a *decision strategy* for designing, implementing and evaluating data-driven decisions, strategies, and actions; (5) to participate in the *leadership* for planning, overseeing and governing hierarchical (e.g., from enterprise-level to groups and teams in the enterprise) visions, missions, plans and strategies; and (6) to manage the *administration* of data, resources, tasks, processes, and risk etc.

Fig. 1 shows the various roles and positions that are needed to undertake the aforementioned areas of responsibility in a corporate data science group [6]. Typical data science roles that differ from existing business intelligence professional roles include data planning strategists, analytical architects, data modelers, subject-specific analysts (e.g., behavior analysts, financial analysts, and visual analysts), discovery scientists, model operators, quality assurers, decision strategists, etc. Interested readers are encouraged to refer to references such as [6] for a more detailed discussion on the various data science positions available, and in particular Chapter 10 Data Professions in [6] for a discussion on the responsibilities

associated with each of the aforementioned areas and the roles in Fig. 1.

B. Data Scientists and Engineers

Of the various data roles, two of the most prominent are data scientists and data engineers. These constitute a collaborative and functional data science team to enable data science and data engineering tasks to be undertaken in an enterprise.

There has been intensive discussion on the roles and responsibilities of data scientists [6], [9], which are usually either broad-based or specific. So, who are data scientists and data engineers? *Data scientists* are data professionals whose major responsibilities and agenda are centered on data-driven exploration and discovery whereas *data engineers* focus on preparing and processing data and the related supporting facilities for data-driven discovery and for obtaining data-driven discovery results. In short, data scientists make sense of data for discovering data value and insights, while data engineers prepare data, support the discovery, and implement the methods and tools to discover data value and insight. Fig. 2 summarizes the responsibilities and qualifications of data scientists and data engineers.

What do data scientists and data engineers do respectively? The main responsibilities of and tasks conducted by data scientists consist of: (1) understanding problem complexities; (2) identifying and specifying constraints and requirements; (3) understanding and quantifying data characteristics; (4) translating underlying challenges to analytical problems; (5) planning analytical strategies and design; (6) conducting data exploration and discovery; (7) evaluating and optimizing analytical results; (8) extracting data value and insight; (9) communicating and interpreting results with stakeholders; and (10) operationalizing data exploration and discovery.

In contrast, data engineers are responsible for (1) understanding the business domain and problems; (2) conducting business requirement analysis; (3) preparing data; (4) preparing the data system; (5) ensuring data quality and ethical compliance; (6) programming data models; (7) computing data; (8) manipulating analytical results; (9) managing projects; and (10) testing data systems and ensuring system quality to achieve the design objectives. In practice, much of this work may be shared and collaboratively undertaken by both data scientists and engineers.

What qualifications should data scientists and data engineers have to fulfil the aforementioned responsibilities and tasks? The following qualifications and capabilities may be required for data scientists: (1) data science thinking; (2) scientific leadership; (3) doctoral qualifications in related fields; (4) capability to handle complex systems and problems-solving; (5) solid foundation in statistics, analytics and learning; (6) hands-on experience in computing; (7) good collaborative, organizational and communication skills; and (8) rich inter-disciplinary knowledge and cross-domain experience.

Data engineers may hold the following qualifications and possess the following skills: (1) knowledge of the methodologies of computing and engineering; (2) master's or doctoral qualifications in related fields; (3) knowledge of applied statistics; (4) knowledge of and skills in software engineering, information systems, programming, distributed and cloud computing, high-performance computing, networking and communication, information retrieval, information security, and enterprise application integration, etc.; (5) skills and practical experience in data processing; (6) experience of system design, implementation, testing, and deployment; (7) knowledge of and experience in project management; (8) skills and experience in user interface design and decision-support systems; and (9) good collaborative, organizational and communication skills.

II. Data Science Competencies

The competencies required by data scientists to conduct creative data science research and actionable practice include the ability to engage in data science thinking and the possession of a data science knowledge base and skill set. The level of maturity of these competencies possessed by either an individual or an organization determines how competitive they are compared to others and how deeply they can delve into data science.

A. Data Science Thinking

What makes data science a new science? There are many possible technical answers to this important question, however the most prevalent is the ability to engage in data science thinking [6]. The critical foundation and success factor of data science is "what and how to think about data", where thinking goes beyond creative and critical thinking and data analytical thinking. *Data science thinking* refers to cognitive and methodological perspectives, thinking traits and habits, and design paradigms and strategies of the mind in handling data problems and systems.

First, from cognitive and methodological perspectives, data science is a higher-level scientific field, a trans-disciplinary science, a complex system, and a comprehensive cognitive and discovery process (for a more specific discussion on these arguments, refer to [6]). Accordingly, a systematic view of scientific methodologies, disciplinary structure, problem complexities, and problem-solving thinking and technical approaches is essential. Integrative systematism [11] may be required to synthesize bottom-up reductionism and top-down holism to generate a systematic view and enable a systematological process of data science. The trans-disciplinary data science synthesizes and transforms multiple relevant sciences and fields such as informatics, computing, statistics and sociology to a new field.

Accordingly, a comprehensive view and structure of data science can be created, such as that shown in Fig. 3 [6]. *Data science thinking* views that a data problem and system is composed of four interconnected and progressive layers: (1) the feed layer, (2) the mechanism design layer, (3) the deliverable layer, and (4) the assurance layer; and two enablers: (1) the thought wing, and (2) the custody wing [6]. This comprehensive understanding of data science starts

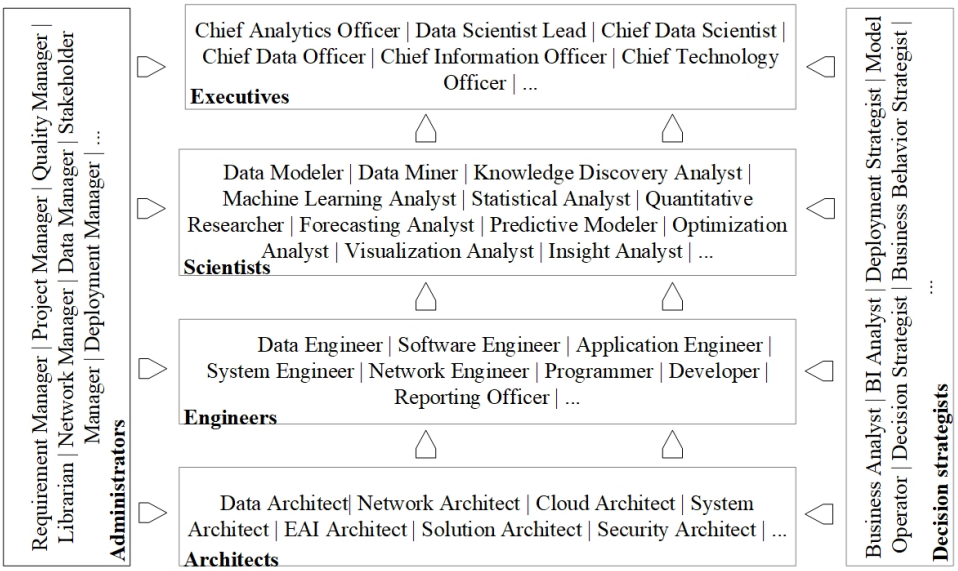

Fig. 1. Data roles for enterprise data science.

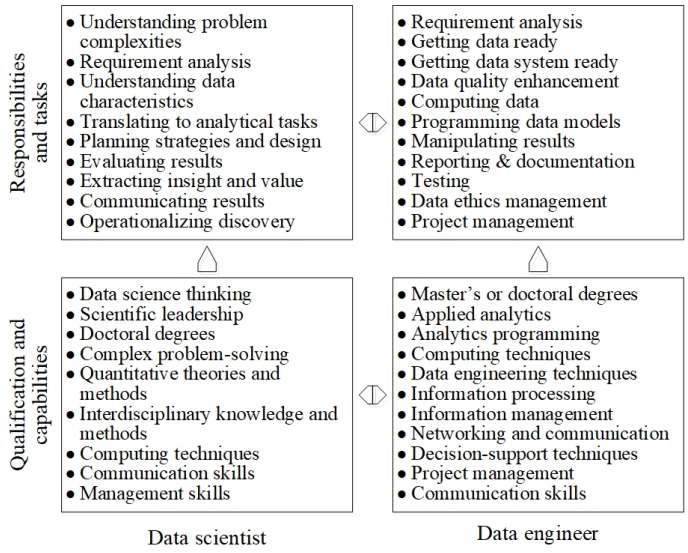

Fig. 2. Responsibilities and qualifications of data scientists and engineers.

with the identification, quantification, and formalization of system complexities of data problems and systems, which involves X-complexities and X-intelligence w.r.t. data, domain, organization, society, human, network and behavior [12], [6].

Second, the aforementioned cognitive and methodological understanding of data science has to be enabled by general and specific data-oriented thinking traits and habits. The exploration of these X-complexities and X-intelligence and their meta-synthesis [11] requires the establishment of scientific thinking for data science, which is complexity and intelligence-driven, data-centric, exploratory and analytical, evidence and fact-based, and reproducible and interpretable. Many critical, creative, and systematic thinking traits and skills are required to achieve data science thinking, e.g., inquisitiveness and imaginative thinking about the *unknown world* [12] w.r.t. unknown data challenges, problems, complexities, hierarchy, structures, distributions, relations, heterogeneities, and unknown capabilities, opportunities and solutions; and the integrity, soundness and interpretability of data-driven exploration processes, analytical and learning theories and methods, and the resultant evidence.

Lastly, data-centric design paradigms, strategies and patterns are necessary to produce original, interesting and actionable [13] data products [6]. One of the most important agendas of the relevant data science communities is to invent appropriate design paradigms, strategies, patterns, architectures, and models to address common or specific data characteristics, system complexities, and outcome expectations. This has been

reflected in the journey of statistical analysis, machine learning, knowledge discovery, and broad informatics and analytics, as exemplified by classic design strategies and patterns such as complementation vs. contrast, individual vs. hybridization design, coupling vs. disentanglement, breadth vs. depth, and structuring (e.g., partition, graph, tree, and ensemble committee); and more recent effort on deepening the network depth to creat deep neural networks and incorporating design strategies such as attention, memory, gating, and adversaries into deep networks. Such effort continues in creating new and more effective and efficient design patterns and strategies to tackle more sophisticated data characteristics and complexities.

B. Data Science Knowledge and Skills

The different data roles, including those of data scientists and data engineers, are empowered by respective bodies of knowledge and the set of skills and capabilities. Since data roles are comprehensive (extending over the whole spectrum of making sense of data and delivering data value) and specific (each role is responsible for particular tasks and expectations), the relevant knowledge map and skill set are broad and can be categorized in terms of different criteria and purposes. Fig. 4 illustrates a framework of competencies for data science teams and individual data scientists [6]. The competency framework suggests data science education and training to foster a comprehensive knowledge and capability set for qualified all-rounded data scientists, including key knowledge, skills and experience in data science thinking, foundational theories, engineering techniques, work-ready practice, communication skills, management, and leadership.

First, it is important to train all those holding a data role to have the ability to engage in *data science thinking* and to develop their mental capabilities so they are able to think about data [14]. The required methodologies, research methods, cognitive skills and mental traits are important to (1) enable general cognitive and scientific thinking, including creative and critical thinking, induction and reduction, abstraction and summarization, logical and imaginative thinking; and (2) develop specific data science thinking traits and methods, including statistical thinking, computational thinking, data analytical thinking, learning and inferential thinking, and optimization and ethics.

Second, the *foundational theories* for data science may involve different disciplines and areas, typically including (1) theories in statistics, mathematics, and complex systems to understand problem complexities and data characteristics; (2) theories and methods for the representation, exploration, analysis, learning and modeling of complex data, behaviors and problems; (3) theories and methods to quantify similarity, dissimilarity, quality, impact and risk; and (4) theories to evaluate and optimize learnability and computational complexity.

Further, many *engineering techniques* are involved in data science and engineering. These include (1) qualitative and quantitative analytical and learning techniques, typically including statistical analysis, knowledge discovery, machine intelligence learning, pattern recognition, natural language processing, and multimedia data analysis, to build and optimize analytical and computational algorithms and models for descriptive, diagnostic, predictive and prescriptive analytics and learning; (2) information processing and management techniques and tools for data preparation, exploratory analysis, information retrieval, data engineering, and analytics programming; and (3) system design, infrastructure building, and decision-supporting techniques and tools to enable human-machine interaction, data infrastructure and platform construction, large-scale computing (including high performance analytics, distributed analytics and cloud analytics), networking and communication, visualization and presentation.

In addition, *work-ready practices* are critical for converting data and data science to extract greater value from data and enhance productivity by applying the aforementioned theories and techniques. Better practices are built on practical skills and hands-on experience of undertaking impactful enterprise data science innovation, actionable and interpretable experimental designs, effective and reproducible analytics projects, exemplary and transferable case studies, and successful widespread applications.

Lastly, the success of data science depends not only on original theories, innovative techniques, and actionable practices, but also on effective communications, management, and leadership. Communication skills are critical in building a high-performing data science team and delivering data science-driven strategic value and decision-making action. *Data science communications* may involve regular workshops and seminars; professional reporting, documentation and visualization of data science design, modeling, processes and results; informative story-telling and evaluation, and business-friendly interpretation of data-driven discovery findings; and team-based collaboration and reflection.

Data science management involves techniques and capabilities for conducting effective data organization, governance and auditing; management of data, resources, projects, roles and responsibilities, and results; assurance of mitigating privacy, security, risk, impact, and social and ethical issues; and optimization of operations, deployment and decisions. In contrast, those who play a *data science leadership* role need knowledge and experience in enterprise data science thinking, research and innovation leadership, strategic planning, decision science, business evaluation, risk management, best practice, and optimization, etc.

C. Data Science Competency Maturity

In reality, the data science competencies between individuals and between organizations are usually divided. The competency imbalance reflects the differences and gaps between organizations which undertake data science.

Data science competency maturity refers to the level of an individual's or organization's capability and capacity to undertake the best possible data science research and best practice. A data science competency maturity model builds benchmarks to structure and quantify the maturity level of competency, which can be measured in terms of data maturity, data science

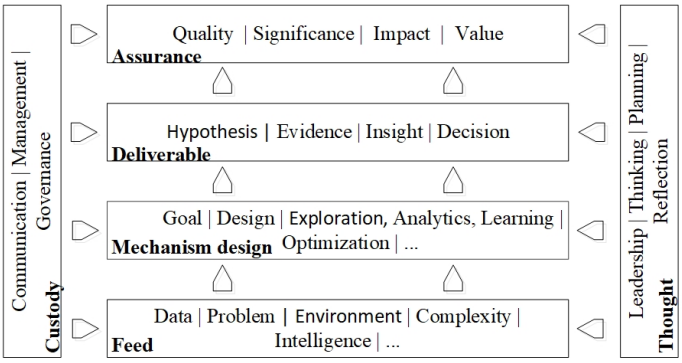

Fig. 3. A data science view and structure.

capability maturity, and data science organizational maturity [6]. Fig. 5 summarizes the various aspects of competency maturity in data science.

Data maturity measures the level of data complexity, the quality of data, the usability of data, the satisfaction rate of data in achieving the anticipated objectives, and the potential of data value. These aspects are specified and quantified in terms of the target sources of data, which may be different, e.g., in some corporate organizations, partial data may be highly mature while the global set may not be.

Data science capability maturity refers to the level of knowledge, capability, practice, and experience of either an individual holding a data role or a data science team in relation to manipulating data and delivering the best possible outcomes. Capability maturity can be structured and quantified in terms of the sufficiency and power level of the essential capabilities, the usability of the capabilities, the level of fit between the capabilities held and those required for utilizing data maturity, achieving data science objectives, and maximizing the value potential of data and capabilities.

Lastly, *data science organizational maturity* refers to the level of organizational maturity in undertaking data science research, innovation, and practice. This is further depicted and categorized in terms of the maturity of strategic organizational data science thinking (thinking maturity), the maturity of data science strategic planning and policies (strategy maturity), organizational infrastructures and capabilities for fulfilling data science (capability maturity), the maturity of organizational data (data maturity), the maturity of organizational data science team (people maturity), and the level of practical experience (practice maturity), and the excellence level of data science organizational leadership (leadership maturity), etc. In addition, data science organizational maturity can be further categorized in terms of organizational hierarchy, e.g., the corporate, departments or business units, teams, and individuals.

III. Data Science Education

There are a large number of courses available online and in academic and training institutions which teach basic to advanced subjects. However, their quality varies significantly and few provide the systematic and intrinsic strategies and programs to train next-generation data scientists and to meet the strategic needs of the booming data economy and data science profession. Here, we briefly review the gaps, and introduce a framework for data science qualifications and pathways to train qualified data professionals.

A. The Gaps in Data Science Education

In recent years, hundreds of courses and subjects have become available at academic institutions and training organizations [9]. These courses are offered in core disciplines of statistics and mathematics, computer science, business and management, and non-core disciplines such as environmental science and finance. It has been observed that almost every major university has either introduced or is about to introduce courses in data science. In addition, online courses, training courses and open courses are increasingly diversified and are attracting teaching resources from traditional academic institutions.

The body of knowledge delivered in typical on-campus data science courses include major subjects such as (applied) statistics, artificial intelligence, machine learning, pattern recognition, big data analytics, and data mining; and minor ones such as programming, visualization, business analytics, project management, and capstone projects. These subjects may be offered at both undergraduate and postgraduate levels.

A comprehensive review of the relevant courses and subjects offered for data science has resulted in the following observations and gaps [6]: (1) none to few courses teach students how to think about data, i.e., the skills and capabilities of data science thinking; (2) truly trans-disciplinary data science courses are few with most courses forming a 'miscellaneous decoupled platter' of existing subjects delivered by different faculties; (3) the level of course quality varies greatly, as does the level of knowledge advancement and the coverage and sufficiency of data science competencies; (4) independent data-driven problem-solving research and innovation capabilities are missing yet critical for conducting real-life data science; (5) there are serious gaps between the challenges of real-world problem complexities and the advancement of knowledge taught in most of the available courses.

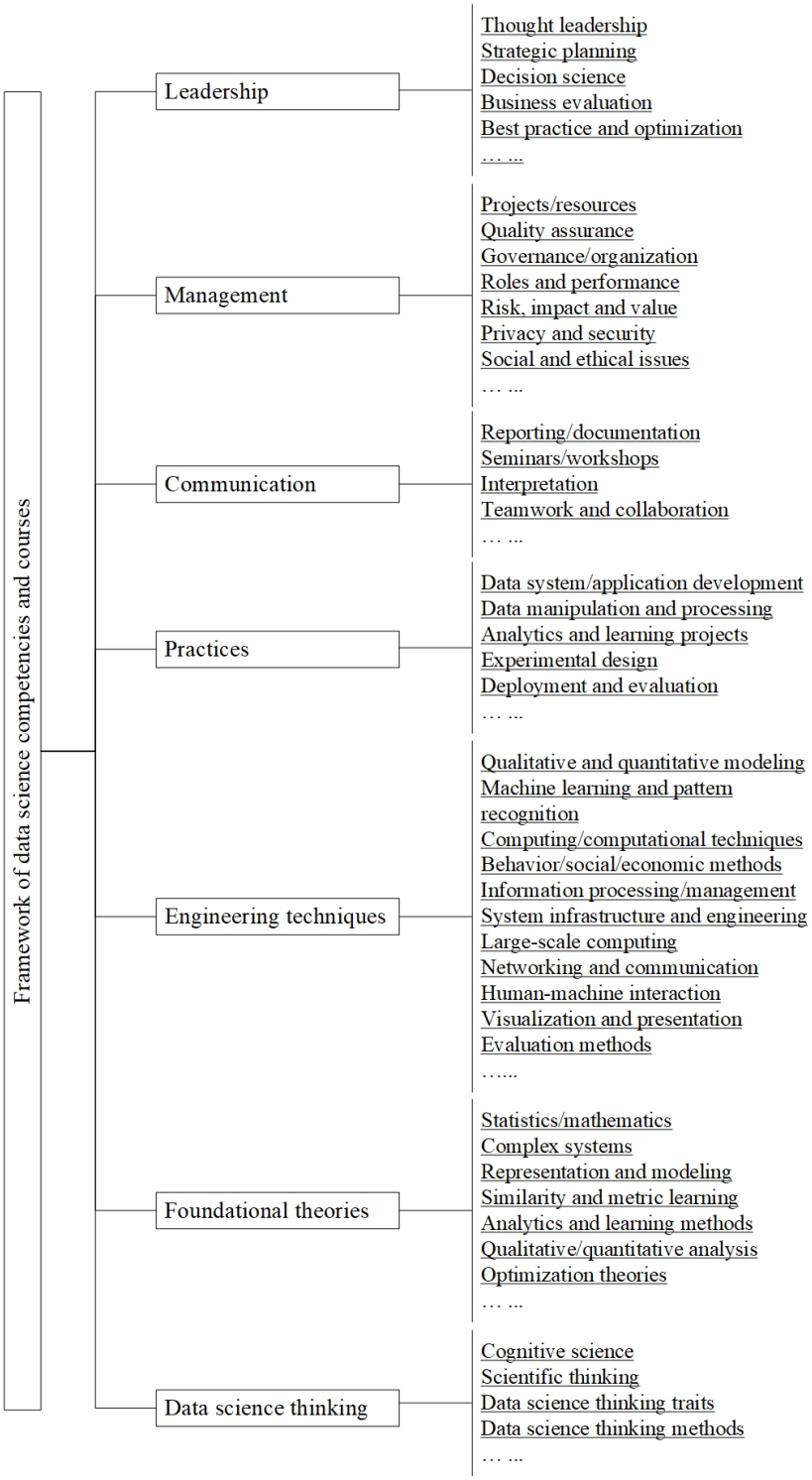

Fig. 4. Framework of data science competencies and courses.

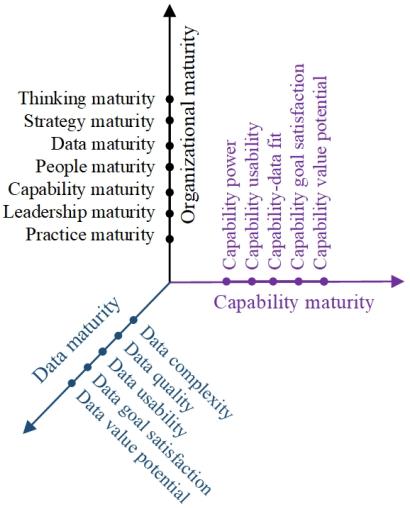

Fig. 5. Data science competency maturity.

B. A Data Science Qualification Framework

Some major questions to be asked when constructing a pragmatic data science course framework include: (1) What makes a qualified next-generation data scientist? (2) What are the gaps and issues in existing data science courses? (3) What is available to foster next-generation data scientists? and (4) How to enable students to cope with unknown challenges and unavailable knowledge and invent the new knowledge required?

It has to be a joint effort between relevant governments, industry, academic institutions, and individuals to systematically address the aforementioned challenges. Various resources and efforts from academia, industry (including corporate training and public courses), and policy-making bodies have to be committed. Here, we focus on how to transform and structure data science courses since they play a driving role in training scientists and engineers. In Fig. 6, a framework of data science qualifications which paves a pathway from undergraduate to master's and doctoral studies is presented.

A *Bachelor in Data Science* builds knowledge, capabilities and experience through a three-stage training process. Stage 1 imparts data science fundamentals which empowers students to engage with data science thinking and provides knowledge on introductory data science and foundations of data science. Stage 2 imparts knowledge and engineering skills to students to enable them to undertake data processing, engineering, computing, and programming, as well as exploratory discovery. Stage 3 focuses on data science practices, which involve multiple capstone projects, internships and applications of data science which enhance the students' skills and experience to ensure sound data understanding, and also covers project management, managing social and ethical issues, and communicating actionable results to stakeholders. As a result, graduates are ready for the workplace in enterprise data science as data analysts or business analysts.

Further, a *Master in Data Science* trains students to become data specialists in a specific business domain or a research area of data science through a three-stage training process. In Stage 1, students obtain advanced foundations in terms of creative data science thinking and foundations for advanced data science. In Stage 2, students utilize the knowledge they have obtained in relation to the advances in data science research and development so they can grasp the latest and most advanced techniques, and are able to conduct advanced data discovery; students are also empowered with the relevant interdisciplinary theories and methods. In Stage 3, students obtain with advanced data science innovation capabilities, enterprise data science innovation and practices, and knowledge and skills of management and leadership.

Lastly, a *PhD in Data Science* imparts the skills required by a senior data science leader to create profound and rigorous theoretical breakthroughs, and provides an in-depth understanding of sophisticated and previously unknown data and problem complexities and the ability to process this. Accordingly, doctoral students focus on developing original and creative philosophical thinking in data science, transforming and creating scientific paradigms for data science research and innovation, and performing novel and significant interdisciplinary research and innovation. They are expected to produce unique and significant research leadership and knowledge advancements and demonstrate better practice in relation to translating high-quality data science advancements into high-impact demonstrations.

IV. Critical Agenda

Data science has been driving new productivity in terms of enabling and creating new science, development, applications, and the economy [6]. Accordingly, the data science profession has emerged as a new profession in which data scientists play profound roles and data science education needs to train the next-generation data scientists to achieve the aforementioned agenda.

The systematic and quality development of the data science profession and education relies on setting up and executing a significant professional and national science and innovation agenda, for example

- conducting comprehensive market surveys of market demand and supply, the quality of and satisfaction with the existing generation of data roles, and the gaps between them and the rapidly developing market need especially that of major data-based innovators, vendors and enterprise end users.
- conducting professional benchmarking and accreditation of data science qualifications and courses that conform to the joint agreement with the relevant disciplines, industry, professional bodies, and policy-making institutions.
- encouraging regional and global collaborations in developing standard guidance and curricula outlines that minimize deviation and maximize the positive experience between institutions, countries and regions.
- implementing educational and training initiatives and programs that support trans-curricula programs to train data

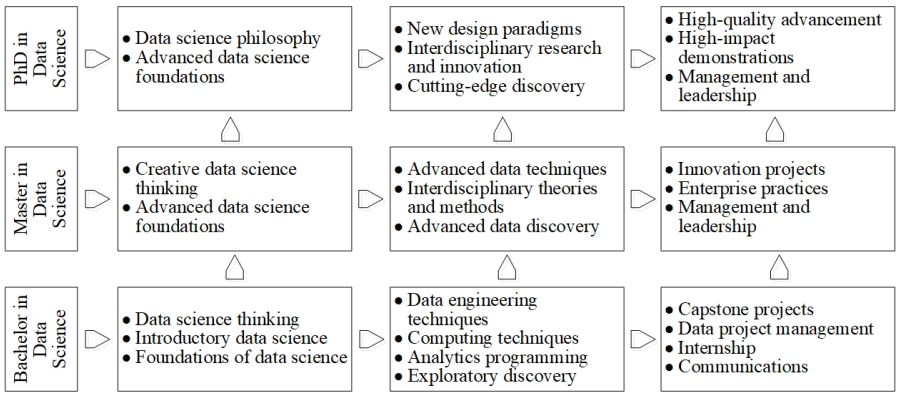

Fig. 6. A framework for data science qualifications and pathways.

scientists with compound trans-disciplinary knowledge and cross-domain experience.

- implementing new educational and training strategies, plans, and programs that substantially enhance the qualifications and capabilities of the existing professionals and train high-calibre next-generation data scientists.

ACKNOWLEDGMENT

This work is partially sponsored by the Australian Research Council Discovery Grant (DP190101079).

REFERENCES

[1] D. J. Trump, "Executive order on maintaining american leadership in artificial intelligence," Feb 2019. [Online]. Available: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[2] The State Council, The People's Republication of China, "Notice of the state council on issuing the development plan on the new generation of artificial intelligence (in chinese)," 7 2017. [Online]. Available: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[3] European Commission, "Artificial intelligence for europe," Apr 2018. [Online]. Available: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625

[4] T. H. Davenport and D. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Business Review*, pp. 70–76, 2012.

[5] L. Cao, "Data science: Nature and pitfalls," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 66–75, 2016.

[6] ——, *Data Science Thinking: The Next Scientific, Technological and Economic Revolution*, ser. Data Analytics. Springer International Publishing, 2018.

[7] R. D. D. Veaux, M. Agarwal, M. Averett, B. S. Baumer, A. Bray, T. C. Bressoud, L. Bryant, L. Z. Cheng, A. Francis, R. Gould, A. Y. Kim, M. Kretchmar, Q. Lu, A. Moskol, D. Nolan, R. Pelayo, S. Raleigh, R. J. Sethi, M. Sondjaja, N. Tiruviluamala, P. X. Uhlig, T. M. Washington, C. L. Wesley, D. White, and P. Ye, "Curriculum guidelines for undergraduate programs in data science," *Annu. Rev. Stat. Appl.*, vol. 4, no. 2, pp. 1–16, 2017.

[8] M. A. Walker, "The professionalisation of data science," *Int. J. of Data Science*, vol. 1, no. 1, pp. 7–16, 2015.

[9] L. Cao, "Data science: A comprehensive overview," *ACM Computing Survey*, vol. 50, pp. 43:1–42, 2017.

[10] UK Government, "Digital, data and technology profession capability framework," Dec 2018. [Online]. Available: https://www.gov.uk/government/collections/digital-data-and-technology-profession-capability-framework

[11] L. Cao, *Metasynthetic Computing and Engineering of Complex Systems*. Springer, 2015.

[12] ——, "Data science: challenges and directions," *Communications of the ACM*, vol. 60, no. 8, pp. 59–68, 2017.

[13] L. Cao, P. S. Yu, C. Zhang, and Y. Zhao, *Domain Driven Data Mining*. Springer, 2010.

[14] B. Baumer, "A data science course for undergraduates: Thinking with data," *The American Statistician*, vol. 69, no. 4, pp. 334–342, 2015.