Computing Competencies for Undergraduate Data Science Curricula

Initial Draft

January 2019

ACM Data Science Task Force

We welcome your feedback. Please submit comments at
https://goo.gl/forms/pCQroVdI8sOtscRi1
by March 31, 2019.

ACM Data Science Task Force

Andrea Danyluk, Co-Chair, Williams College & Northeastern University, USA

Paul Leidig, Co-Chair, Grand Valley State University, USA

Scott Buck, Intel Corporation, USA

Lillian Cassel, Villanova University, USA

Andrew McGettrick, University of Strathclyde, UK

Weining Qian, East China Normal University, China

Christian Servin, El Paso Community College, USA

Hongzhi Wang, Harbin Institute of Technology, China

CONTENTS

Chapter 1: Introduction

1.1 Charter

At the August 2017 ACM Education Council meeting, a task force was formed to explore a process to add to the broad, interdisciplinary conversation on data science, with an articulation of the role of computing discipline-specific contributions to this emerging field. Specifically, the task force would seek to define what the computing/computational contributions are to this new field, and provide guidance on computing-specific competencies in data science for departments offering such programs of study at the undergraduate level.

There are many stakeholders in the discussion of data science – these include colleges and universities that (hope to) offer data science programs, employers who hope to hire a workforce with knowledge and experience in data science, as well as individuals and professional societies representing the fields of computing, statistics, machine learning, computational biology, computational social sciences, digital humanities, and others. There is a shared desire to form a broad interdisciplinary definition of data science and to develop curriculum guidance for degree programs in data science.

This volume builds upon the important work of other groups who have published guidelines for data science education. There is a need to acknowledge the definition and description of the individual contributions to this interdisciplinary field. For instance, those interested in the business context for these concepts generally use the term “analytics”; in some cases, the abbreviation DSA appears, meaning Data Science and Analytics.

This volume is a draft articulation of computing-focused competencies for data science. It recognizes the inherent interdisciplinarity of data science and situates computing-specific competencies within the broader interdisciplinary space.

1.2 Prior work on defining data science curricula

As an inherently interdisciplinary area, data science generates interest within many fields. (See Figure 1.) Accordingly, there have been a number of Data Science curriculum efforts, each reflecting the perspective of the organization that created it.

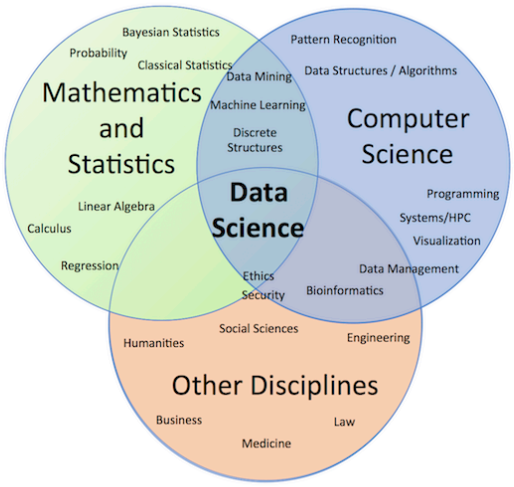


Figure 1

This project looks at data science from the perspective of the computing disciplines, but recognizes that other views contribute to the full picture. The following examples are especially important, and have informed the committee's work.

The EDISON Data Science Framework (2018)

EDISON is a project started in September 2015 "with the purpose of accelerating the creation of the Data Science profession." The core EDISON consortium consists of seven partners across Europe. Since 2015, the group has worked to create the EDISON Data Science Framework. This collection of documents includes a general introduction, as well as four detailed components, including:

- Data Science Competences Framework
- Data Science Body of Knowledge
- Data Science Model Curriculum
- Data Science Professional Framework

This comprehensive set of curricular volumes parallels the intended structure of our work. EDISON was in earlier stages as this project began; at present, it is clear that there are significant overlaps, and future versions of our work with reconcile our model with the EDISON

curriculum, with the intention of creating a complementary volume, rather than a replicated or competing volume.

The National Academies of Science, Engineering, and Medicine Report on Data Science for Undergraduates (2018)

As the press release announcing the publication of the National Academies report states, "Data science draws on skills and concepts from a wide array of disciplines that may not always overlap, making it a truly interdisciplinary field. Students in many fields need to learn about data collection, storage, integration, analysis, inference, communication, and ethics." The report highlights the demand for data scientists and calls for a broad education for students across programs of study. Identifying many data science roles, including those related to hardware and software platforms, data storage and access, statistical modelling and machine learning, and business analytics, among others, the report does not presume that every data scientist will be expert in all areas, but rather that programs will develop to allow graduates to fulfil specific roles.

The intent of the National Academies report was to highlight the importance, breadth, and depth of data science, and to provide high-level guidance for data science programs. It is not a detailed curricular volume in the sense of the EDISON project or this ACM Data Science effort.

The Park City Report (2017)

The Park City Math Institute 2016 Summer Undergraduate Faculty Program convened with the purpose of articulating guidelines for undergraduate programs in data science. The three-week workshop brought together 25 faculty from computer science, statistics and mathematics. The base assertion of the report and proposed curriculum is that data is the core: "The recursive data cycle of obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses, and communicating the results lies at the core of the data science experience."

The resulting list of key competencies shows the interdisciplinary nature of data science, with an understandable focus on the mathematics and statistics:

- Computational and statistical thinking
- Mathematical foundations
- Model building and assessment
- Algorithms and software foundation
- Data curation
- Knowledge transference – communication and responsibility

The role of computer science appears in the description of computational thinking: "Data science graduates should be proficient in many of the foundational software skills and the associated algorithmic, computational problem solving of the discipline of computer science." However, further description relates these skills to understanding the programming and algorithms behind "professional statistical analysis software tools."

The Park City report deserves further description. It includes an outline of the Data Science Major:

1. Introduction to data science
    a. Introduction to Data Science I
    b. Introduction to Data Science II
2. Mathematical foundations
    a. Mathematics for Data Science I
    b. Mathematics for Data Science II
3. Computational thinking
    a. Algorithms and Software Foundations
    b. Data Curation—Databases and Data Management
4. Statistical thinking
    a. Introduction to Statistical Models
    b. Statistical and Machine Learning
5. Course in an outside discipline

The report also includes a description of each of the courses. For the purposes of this report, it is noted that programming is introduced in Introduction to Data Science I and II, and appears again as a part of Algorithms and Software Foundations. The course in Data Curation includes traditional databases as well as newer approaches to data storage and interaction. The course in Statistical and Machine Learning "blends the algorithmic perspective of machine learning in computer science and the predictive perspective of statistical thinking."

Although there certainly are additional aspects of computer science that are relevant to the preparation of a student of data science, there is clearly an effort to combine the mathematical and computer science contributions to produce a blended program. This ACM Data Science report builds on the Park City work with a heavy orientation toward computer science. The position of the Task Force is that any Data Science program will have to reflect competencies in mathematics, statistics, and computer science, possibly with different emphases. This is consistent with the view of the National Academies report. Graduates of programs following the Park City guidelines will have valuable strengths and graduates of programs following these ACM guidelines will have different, but equally valuable strengths.

The Business Higher Education Framework (BHEF) Data Science and Analytics (DSA) Competency Map (2016)

The work provides a four-level competency map. The base, or Tier 1, level describes personal effectiveness competencies. These are not considered competencies learned in school, but rather part of an individual's personal development. Examples include integrity, initiative, dependability, adaptability, professionalism, teamwork, interpersonal communication, and respect.

Tier 2 describes academic competencies to be acquired in higher education. These are most relevant to this report and include the following:

- Deriving value from data
- Data literacy
- Data Governance and Ethics
- Technology
- Programming and Data Management
- Analytic Planning
- Analytics
- Communication

Tier 3 presents workplace competencies: planning and organizing, problem solving, decision-making, business fundamentals, customer focus, and working with tools and technology.

Tier 4 is for Industry-Wide Technical Competencies. These are not specified, but represent skills that are common across sectors of a larger industry context.

Though Tier 2 includes a competency in "Programming and Data Management," the description mentions only "Write data analysis code using modern statistical software (e.g., R, Python, and SAS)." This set of competencies does not address a need for developing new software or systems in support of data science, but relies on available tools.

Business Analytics Curriculum for Undergraduate Majors (2015)

This report was produced in 2015 by the Institute for Operations Research and the Management Sciences (INFORMS). Reflecting the focus of programs in Business, this INFORMS curriculum assumes basic computer literacy as a starting point. It suggests revising some of the standard courses in statistics to meet newer needs. The resulting course list includes: Data Management, Descriptive Analytics, Data Visualization, Predictive Analytics, Prescriptive Analytics, Data Mining, and Analytics Practicum. It also includes electives.

Like the guidelines from the Business Higher Education Framework, the focus is on doing something with data, primarily to serve business needs. There is no mention of programming. The data management course includes SQL, but has no prerequisites. The emphasis in the data mining course is on framing a business problem. Data mining techniques are compared, and large datasets are to be used. The tools to be used for that purpose are not specified.

Initial workshops related to this ACM Data Science Curriculum effort (2015)

In October 2015, the National Science Foundation sponsored a workshop with representatives of many perspectives on data science. Some attendees represented established programs, others represented societies with an interest in data science. The final report, "Strengthening Data Science Education Through Collaboration," describes the discussions and reflects the diversity of opinions. Although opinions varied, there were some areas of agreement. Those form the basis of the list of Knowledge Areas in this current ACM report.

Summary

The review of existing curricular efforts suggests that it would be important to capture in a single volume the contributions that computing makes to data science. Through developments such as the Internet of Things, sophisticated sensors, face recognition and voice recognition, automation, etc., computing opens up many avenues for data collection. It can play a vital role as a custodian of information with great attention being paid to maintenance but crucially also to security and confidentiality matters. Then the analysis of large amounts of information and utilization of that for the purposes of machine learning or augmented intelligence in its various roles can bring significant benefit.

1.3 Committee Work and Processes

The Data Science Task Force was initiated at a meeting of the ACM Education Council in August 2017. The Co-Chairs were appointed at the meeting and were charged with developing a charter for the work, as well as assembling a task force with global representation.

The Co-Chairs drafted a proposal to create the Task Force, which was approved by the ACM Education Board in January 2018. The initial Task Force – approximately two-thirds of the members of the current committee – convened for a full-day meeting in February 2018.

In preparation for a second face-to-face meeting in July 2018, the Task Force designed two surveys to gather input from academia and industry on the computing competencies most central to Data Science. The results of the survey are presented in this report, with details provided in Appendix B. During this time, the Co-Chairs invited additional members to join the committee and began to develop a global advisory group.

At the July 2018 meeting, the ACM Task Force developed the set of computing-focused Knowledge Areas for Data Science that appear in this report and began to articulate competencies in each of those areas.

With the release of this first draft report, the ACM Data Science Task Force is calling for discussion and feedback from all data science constituencies. The Task Force will be presenting the report and gathering comments at conferences and meetings, including Educational Advances in Artificial Intelligence (EAAI-9), held at AAAI in January 2019; the SIGCSE Symposium in February 2019; and the Joint Statistical Meetings in July 2019. The Task Force also welcomes feedback by email to the Co-Chairs:

- Andrea Danyluk (andrea@cs.williams.edu)
- Paul Leidig (leidigp@gvsu.edu)

1.4 Survey of Academic and Industry Representatives

In order to gain an understanding of the current data science landscape, the ACM Data Science Task Force conducted a survey of ACM members, representing academic institutions and industry organizations. Through outreach to ACM members, the Task Force was also able to reach computing professionals outside of ACM membership. In all cases, the Task Force sought global participation. There were 672 responses to the academic survey and 297 responses to the industry survey.

Academic Survey

The academic survey asked academics whether their institution had any sort of data science program at the undergraduate level, asked what type of program was offered, in what department(s) it was housed, and what computing areas were required, elective, or not present in the program. It also allowed respondents to add to the list of computing areas specified in the survey. Finally, the survey asked participants whether their data science program had a "data science in context" requirement – i.e., a requirement that students apply data science to another area.

Nearly half of respondents from academic institutions (47%) reported they did not offer an undergraduate data science program. However, over half of those who reported offering some type of program offered a full bachelor's degree in data science.

Nearly all of the programs offering a bachelor's degree in data science required courses in programming skills and statistics. In addition, the majority of programs also required data management principles, probability, data structures and algorithms, data visualisation, data mining, and machine learning. Other courses included topics such as ethics, calculus, discrete mathematics and linear algebra. We note that a majority of programs also required a "data science in context" course.

Administratively, the largest percentage of programs were housed in a computer science department; however, almost as many were in an "other" category. This result might be somewhat skewed, given that the survey was fielded primarily with ACM members.

Additionally, over half of these programs reported graduating 10 students or less annually.

We expect that the number of Data Science programs will increase, as will the number of students choosing to study it. This, then, is an ideal time to articulate computing-based competencies for those programs.

Industry Survey

The industry survey roughly mirrored the academic survey; however, the primary question was whether a company looked for job applicants with data science experience and what computing experience they required or preferred those applicants to have.

In the survey of industry representatives, nearly half (48%) responded that they look for candidates specifically with data science or analytics degrees or educational backgrounds. We found it particularly interesting that the majority of employers reported these employees work as individual contributors on data science tasks.

Industry respondents reported requiring experience or skills in similar areas to those required by college or university Data Science programs. One slight difference is that employers reported requiring more computing skills than statistical or mathematical skills.

Other Observations

The ACM Task Force was somewhat surprised by certain survey results. For instance, industry respondents did not report data security and privacy as a required competency area for job applicants. We note that this may reflect employers' understanding of what Data Science (and Computer Science) programs are requiring of their majors. That is, it might reflect the reality of the applicant pool, rather than a "wish list" of competencies.

Similarly, we note that academic institutions reported what they currently require, rather than what they would require in an ideal world. This might, in some cases, reflect the availability of courses and faculty at an institution, rather than a "gold standard" for Data Science programs.

A more detailed summary of survey results is presented in Appendix B.

1.5 Knowledge Areas

Following the work of previous ACM curricular volumes (see [ACM 2013], for instance), this report is organized around Knowledge Areas (KAs) whose origins are based on survey input (see Section 1.4) as well as prior work, with special attention being given to the results of the workshop reported in [CasTopi 2015].

The core computing discipline-specific Knowledge Areas for Data Science are:

- Computing Fundamentals, including Programming, Data Structures, Algorithms, and Software Engineering
- Data Acquirement and Governance
- Data Management, Storage, and Retrieval
- Data Privacy, Security, and Integrity
- Machine Learning
- Data Mining
- Big Data, including Complexity, Distributed Systems, Parallel Computing, and High Performance Computing
- Analysis and Presentation, including Human-Computer Interaction and Visualization
- Professionalism

Other areas of computing may merit attention: sensors and sensor networks, the Internet of Things, vision systems, among others.

In addition, for a full curriculum the above need to be augmented with courses covering calculus, discrete structures, probability theory, elementary statistics, advanced topics in statistics, and linear algebra.

1.6 Data Science in context

In addition to developing foundational skills in computing and statistics, data science students should also learn to apply those skills to real applications. It is important for data science education to incorporate real data used in an appropriate context.

Data Science curricula should include courses designed to promote dual coverage combining both data science fundamentals and applications, exploring why people turn to data to explain contextual phenomena. Such courses highlight how valuable context is in data analytics; where data are viewed with narratives, and questions often arise about ethics and bias. It can be beneficial to teach some courses with a disciplinary context so that students appreciate that data science is not an abstract set of approaches. Related application disciplines might include physics, biology, chemistry, the humanities, or other areas.

1.7 Competency Framework

The Competency Framework provides a framework for the description of the various Knowledge Areas. Each KA is described by some preliminary descriptive material followed by a set of topics and a set of associated competencies; levels of competence vary, with some requiring greater expertise than others.

The details of the Competency Framework are described in Chapter 2. The descriptions of the Knowledge Areas are then provided in Appendix A.

1.8 Motivating the Study of Data Science

Those who study Data Science have to develop a mind set with a strong focus on data – the collection of data and, through analysing it appropriately, using this to bring about beneficial insights and changes. For instance:

- Obtaining data about the quality of air in a city can result in removing dangerous pollution or sending warning messages to those who suffer from asthma.
- Collecting data about traffic in real time can result in steps being taken to avoid traffic congestion.
- Collecting patient data can lead to new insights for disease diagnosis and treatment.
- Recording data about speech in a certain area can assist with speech recognition.

The possibilities are endless, and the contributions that Data Science can make to transforming businesses, transforming society and basically shaping the future for the better are huge. The possibilities also carry with them potentially negative consequences.

Students of Data Science need to be imbued with the 'joy of data', seeing data as the 'currency or fuel of our time'. They also need to be imbued with a strong sense of professional and ethical responsibility. Data Science courses ought to reflect such sentiments; likewise the education of data scientists.

The topic of careers is of course important from a marketing perspective. Suffice it to say that the current demand is considerable and growing daily.

1.9 Overview of this Report

Having set the scene in this chapter, the second chapter sets out the Competency Framework used in describing the various Knowledge Areas in some detail. The computing related KAs are captured in Appendix A.

References

[ACM 2103] *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science (ACM/IEEE 2013):* *https://www.acm.org/education/CS2013-final-report.pdf*

[ASA 2014] *Curriculum Guidelines for Undergraduate Programs in Statistical Science (ASA 2014b): http://www.amstat. org/education/pdfs/guidelines2014-11-15.pdf>*

[BHEF 2106] *Data Science and Analytics (DSA) Competency Map, Business The Business Higher Education Framework (BHEF) version 1.0 produced in* November 2016

[CasselTopi 2015] *Strengthening Data Science through Collaboration,* by Lillian Cassel and Heikki Topi, Technical Report and report of 2015 NSF Workshop. http://www.computingportal.org/sites/default/files/Data%20Science%20Education%20Workshop%20Report%20.0_0.pdf

[CUPM 2015] *Curriculum Guide to Majors in the Mathematical Sciences (MAA 2015).* See http://www.maa.org/ sites/default/files/pdf/CUPM/pdf/CUPMguide_print.pdf

[EDISON] *The Edison Data Science Competence Framework* http://edison-project.eu/edison/edison-data-science-framework-edsf.

[Edison 2015] Data science professional uncovered: How the Edison project will contribute to a widely accepted profile for data scientists, by Manieri, A.; Brewer, S.; Riestra, R.; Demchenko, Y.; Hemmje, M.; Wiktorski, T.; Ferrari, T.; and Frey, J. published in IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), 588–593. National Academies of Sciences, Engineering, and Medicine. 2018.

[INF 2015] *Business Analytics Curriculum for Undergraduate Majors*, Coleen R. Wilder, Ceyhun O. Ozgur (2015) published in INFORMS Transactions on Education 15(2):180-187. https://doi.org/10.1287/ited.2014.0134

[NatAc 2018] *Data Science for Undergraduates: Opportunities and Options*, published by the National Academies of Sciences, Engineering, and Medicine, 2018. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104

[Park City 2017] *Curricular Guidelines for Undergraduate Programs in Data Science* by DeVeaux, R.; Agarwal, M.; Averett, M.; Baumer, B.; Bray, A.; Bressoud, T.; Bryant, L.; Cheng, L.; Francis, A.; Gould, R.; Kim, A.; Kretchmar, M.; Lu, Q.; Moskol, A.; Nolan, D.; Pelayo, R.; Raleigh, S.; Sethi, R.; Sondjaja, M.; Tiruviluamala, N.; Uhlig, P.; Washington, T.; Wesley, C.; White, D.; and Ye, P. 2017. Annual Review of Statistics and Its Application 4:15–30.

Chapter 2: The Competency Framework

Much of the material in this chapter leans very heavily on (i.e., is taken verbatim from) the work of IT2017 – see [IT2017]. The motivation for this is to maintain consistency across a set of curricula documents produced by ACM.

2.1 Competency in theory

2.1.1 Meaning of Competency

Learning outcomes are written statements of what a learner is expected to know and be able to demonstrate at the end of a learning unit (or cohesive set of units, course module, entire course, or full program).

In contrast, with the wide agreement on the meaning of learning outcomes, there is extensive confusion and vagueness around the terms competence and competency. Generally, the term competence refers to the performance standards associated with a profession or membership to a licensing organization. Assessing some level of performance in the workplace is frequently used as a competence measure, which means measuring aspects of the job at which a person is competent. Competencies are what a person brings to a job conceptualized as qualities by which people demonstrate superior job performance [Kli1].

There is general agreement in education that success in college and career readiness requires that students develop a range of qualities [Ken1, Nas1, Nrc1], typically organized along three dimensions: knowledge, skills, and dispositions. We utilize a working definition of *competency* that connects knowledge, skills, and dispositions. Figure 2.1 (adapted from IT2017, which is, in turn, adapted from [Ccs1, p. 5]) shows these interrelated dimensions of competency.

COMPETENCY = KNOWLEDGE + SKILLS + DISPOSITIONS

KNOWLEDGE
• Mastery of content knowledge • Transfer of learning

SKILLS
• Capabilities and strategies for higher-order thinking
• Interactions with others and world around

DISPOSITIONS
• Personal qualities (socio-emotional skills, behaviors, attitudes) associated with success in college and career

Figure 2.1 Interrelated dimensions of competency

In the working definition of competency, the three interrelated dimensions have the following meanings:

- *Knowledge* designates a proficiency in core concepts (or topics) and content of Data Science and application of learning to new situations. This dimension usually gets most of the attention from teachers, when they design their syllabi; from departments, when they develop program curriculum; and from accreditation organizations, when they articulate accreditation criteria.

- *Skills* refer to capabilities and strategies that develop over time, with deliberate practice and through interactions with others and the real world. [Nrc1]. Skills also require engagement in higher-order cognitive activities, meaning that "hands-on" practice of skills join with a "minds-on" engagement. The inextricable connection between knowledge and skills is evident in Michael Polanyi's characterization of explicit versus tacit knowledge [Pol1]. Explicit knowledge, or "know-that," reflects core ideas and principles, and corresponds to the knowledge dimension in our definition. Tacit knowledge, or "know-how," is a skillful action requiring sustained engagement and practice. Problem-based assignments, real-world projects, and laboratory activities with workplace relevance are examples of curriculum elements that focus on developing skills. Well-designed syllabi and accredited programs are mindful of skill development when they articulate student outcomes at course and program level.

- *Dispositions* encompass socio-emotional skills, behaviors, and attitudes that characterize the inclination to carry out tasks and the sensitivity to know when and how to engage in those tasks [Per2]. Originating from the field of vocational education and research on career development, dispositions have received increasing attention in the K–12 computer science education community [Ste1]. Formulating an operational definition of computational thinking, Barr and Stephenson [Bar2] included the dispositions category to capture areas of values, motivation, feelings, stereotypes, and attitudes such as confidence in dealing with complexity, tolerance to ambiguity, persistence in working with difficult problems, and knowing one's strengths and weaknesses and setting aside differences when working with others. To distinguish dispositions from knowledge and skills, we use Schussler's view that a disposition "concerns not *what* abilities people have, but *how* people are disposed to use those abilities." [Sch1]

2.1.2 A Performance Perspective on Learning

A transmission theory of teaching, also known as teacher-focused, holds that knowledge emerges as it transmits from the expert teacher to the inexpert learners with the objective of 'getting it across' or covering all the topics in the material. The opposing theory of active learning is that students themselves create meaning and develop understandings with the help of appropriately designed learning activities. In undergraduate education, the active learning model underlies a shift of the paradigm that has governed higher education institutions. The traditional paradigm of *providing instruction* dominated by a passive lecture-based learning environment has shifted to *producing learning* and creating experiences in which students are active participants in the learning process [Bar1].

On a student learning continuum from passive (attending a standard lecture) to active (engaged in problem solving with peers), to produce high level of student engagement means to design learning activities in which students do more than taking notes, recalling, observing, or describing. They learn more effectively when their active participation consists of asking questions, applying concepts, discovering relationships, or generalizing a solution to new situations [Big2]. Higher level of engagement cannot be encouraged if teaching is only about declarative and procedural knowledge: information, vocabulary, basic concepts, basic knowhow, and discrete skills [Wig1]. Indeed, students need the acquisition of knowledge and development of basic skills, but this is just a means to a more important preparation for authentic performance tasks and transfer of learning in new situations.

Perkins and Blythe formulated a "performance perspective" of learning and offered the view that "understanding something is a matter of being able to carry out a variety of performances concerning the topic." [Bly1, Bly2] A performance perspective of learning requires a "modicum of *transfer*, because it asks the learner to go beyond the information given" and seeks to "... transcend the boundaries of the topic, the discipline, or the classroom." [Per1]

2.1.3 Learning Transfer

The conventional way of framing curriculum guidelines for computing programs, has, until recently, been content driven. A disciplinary body of knowledge decomposes into areas, units, and topics to track recent developments in a rapidly changing computing field. For this report, we follow the approach of the IT2017 report, which used the *Understanding by Design* (UbD) framework [Wig1] to present a competency-based curricular framework.

The idea of the UbD framework is to treat content mastery as a means, not the end, to long-term achievement gains that a program of study envisions for its graduates. Learners could know and do many discrete things, but still not be able to see the bigger picture, put it all together in context, and apply their learning autonomously in new situations.

In the UbD framework, learning transfer is multi-faceted as shown in Table 2.1 [Wig2]. We note that these facets of learning transfer blended skills and dispositions. Explain, interpret, apply and adjust are skills complemented by dispositions related to showing empathy, perceiving sensitively, recognizing bias, considering various points of view, or reflecting on the meaning of new learning and experiences. Dispositions relating to meta-cognitive awareness include being responsible, adaptable, flexible, self-directed, and self-motivated, and having self-confidence, integrity, and self-control. They also include how we work with others to achieve a common goal or solution.

Table 2.1: Six facets of learning transfer adapted from Understanding by Design framework and reproduced from IT2017.

Explain	Learners make connections, draw inferences, express them in their own words with support or justification, use apt analogies; teach others.
Interpret	Learners make sense of, provide a revealing historical or personal dimension to ideas, data, and events; interpretation is personal and accessible through images, anecdotes, analogies, and stories; turn data into information; provide a compelling and coherent theory.
Apply	Learners use what they have learned in varied and unique situations; go beyond the context in which they learned to new units, courses, and situations, beyond the classroom.
Demonstrate Perspective	Learners see the big picture, are aware of, and consider various points of view; take a critical and disinterested stance; recognize and avoid bias in how positions are stated.
Show Empathy	Learners perceive sensitively; can "walk in another's shoes;" find potential value in what others might find odd, alien, or implausible.
Have Self-Knowledge	Learners show meta-cognitive awareness on motivation, confidence, responsibility, and integrity; reflect on the meaning of new learning and experiences; recognize the prejudices, projections, and habits of mind that both shape and impede their own understanding; are aware of what they do not understand in a specific context.

2.2 Competencies and professional practice

On a practical, operational level, competencies are conceptualized as higher-level learning outcomes linked to performance tasks and are descriptive of the professional context of those tasks. We follow the Van der Klink and Boon advice that the "fuzziness" of competencies "disappears in the clarity of learning outcomes." [Kli1] A sensible method to articulate competencies is to select learning outcomes that lead to achieving those competencies along with evaluation indicators suggestive of a professional context [Ken2]. A performance perspective on learning [Per1] is not possible without performance-based assessments. The design of performance assessments considers authentic situations and aspects of work that professionals encounter and through which they demonstrate expertise.

A competency-based approach to a Data Science curricular framework considers the long-term goal of learning to *achieve genuine competence through ongoing transfer* of what students learn and graduates develop in their professions and advanced academic studies. To articulate performance goals for each Data Science domain, the task group follows the recommendation of IT2017: the UbD approach of considering performance verbs associated with the six facets of learning transfer: explain, interpret, apply, demonstrate perspective, show empathy, and have self-knowledge as described in Table 2.1. A sample list of performance verbs that generate ideas

for performance goals and professional practice [Wig2] appears in Table 2.2 below; they are useful in describing the Data Science competencies expected from Data Science graduates.

Table 2.2: Performance verbs to generate ideas for performance goals and professional practice (Reproduced from IT2017)

Explain	**Interpret**	**Apply**	**Demonstrate Perspective**	**Show Empathy**	**Have Self-Knowledge**
demonstrate derive describe how design exhibit express induce instruct justify model predict prove show how synthesize teach	create analogies critique document evaluate illustrate judge make sense of make meaning of provide metaphors read between the lines represent tell a story of translate	adapt build create debug decide design exhibit invent perform produce propose solve test use	analyze argue compare contrast criticize infer	assume role of be like be open to believe consider imagine relate role play	be aware of realize recognize reflect self- assess

References

[Bar1] Barr, R.B. and Tagg, J. 1995. From Teaching to Learning: A New Paradigm for Undergraduate Education. *Change*, 27(5), 12-25.

[Bar2] Barr, V. and Stephenson, C. 2011. Bringing computational thinking to K-12: What is involved and what is the role of computer science education community? *ACM Inroads*, 2, 1 (May 2011), 48-54.

[Big2] Biggs. J. 1999. *Teaching for Quality Learning at University – What the Student Does* (1st Edition), SRHE / Open University Press, Buckingham.

[Bly1] Blythe, T. 1998. *The teaching for understanding guide*. San Francisco: Jossey-Bass. Blythe, T, and Perkins, D. 1988. Understanding understanding. In T.

[Bly2] Blythe (Ed.), *The teaching for understanding guide,* 9-16. San Francisco: Jossey-Bass.

[Ccs1] Council of Chief State School Officers. 2013. *Knowledge, Skills, and Dispositions: The Innovation Lab Network State Framework for College, Career, and Citizenship Readiness, and Implications for State Policy*.

[IT2017] Information Technology 2017, Final Curriculum Report IT2017 published by ACM on 10th December 2017

[Ken1] Kennedy, D., Hyland, Á., & Ryan, N. 2007. *Writing and using learning outcomes: a practical guide*. Cork: University College Cork.

[Ken2] Kennedy, D., Hyland, A, and Ryan, N. 2009. *Learning Outcomes and Competences*. *Bologna Handbook*. *Introducing Bologna Objectives and Tools*, B 2.3-3, 1-18.

[Kli1] Klink M. van der, Boon, J., and Schlusmans, K. 2007. Competences and vocational higher education: Now and in future. *European Journal of Vocational Training* No 40 – 2007/1, 67-82.

[Nas1] National Academies of Sciences, Engineering, and Medicine. 2016. *Supporting Students' College Success; Assessment of Intrapersonal and Interpersonal Competencies*. Washington, DC: The National Academies Press. https://doi.org/10.17726/24697.

[Nrc1] National Research Council. 2012. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13398.

[Per1] Perkins, D. 1993. Teaching for understanding. *American Educator: The Professional Journal of the American Federation of Teachers,* 17(3), 8, 28-35, Fall 1993.

[Per2] Perkins, D., Jay, E., and Tishman, S. 1993. Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly*, 39(1), 1-21.

[Pol1] Polanyi, M. 1966. *The Tacit Dimension*. University of Chicago Press: Chicago.

[Sch1] Schussler, D.L. 2006. Defining dispositions: Wading through murky waters. *The Teacher Educator*, 41(4).

[Ste1] Stephenson, C. and Malyn-Smith, J. 2016. *Computational thinking from a dispositions perspective*. The Keyword, Google Education. Accessed July 6, 2017, https://www.blog.google/topics/education/computational-thinking- dispositions-perspective/?m=0.

[Wig1] Wiggins, G.P., McTighe, J., and Ebrary, I. 2005. *Understanding by design* (Expanded second edition). Alexandria, VA: Association for Supervision and Curriculum Development.

[Wig2] Wiggins, G., and McTighe, J. 2011. *The Understanding by Design Guide to Creating High-Quality Units*. Alexandria, VA: Association for Supervision and Curriculum Development.

Appendix A: A Draft of Competencies for Data Science

Computing Fundamentals

Data scientists should be able to implement and understand algorithms for data collection and analysis. They should understand the time and space considerations of algorithms. They should follow good design principles developing software, understanding the importance of those principles for testability and maintainability.

Programming

This includes development and implementation of algorithms, as well as integration with existing software and/or tools.

Computing Fundamentals: Programming	
Scope ● Development and implementation of algorithms, including integration with various existing software and/or tools. ● Usage of traditional programming languages to integrate existing interfaces between datasets and applications. ● Usage of a programming language designed for statistical computing in the context of a data science problem. ● Abstract Data Types (ADT) to create a simple program ● Potential vulnerabilities by a given program ● Programming code that utilizes preconditions, postconditions, and invariants.	*Competencies* ● Design an algorithm in a programming language to solve a simple problem ● Use the techniques of decomposition to modularize a program. ● Create code in a programming language that includes primitive data types, references, variables, expressions, assignments, I/O, control structures, and functions ● Create a simple program that uses recursion. ● Illustrate the use of databases and apply SQL and NoSQL ● Write a regular expression to match a pattern ● Use standard libraries for a given programming language ● Design and implement programs that use a database ● Use techniques for searching patterns in data ● Implement good documentation practices in programming

Data Structures

A data scientist should know a variety of data structures, be able to use them, and understand the implications of choosing one over another.

Computing Fundamentals: Data Structures	
Scope ● Classification of data storage, accessibility and complexities, based on implementation and operation of domain-cluster problems and/or applications ● Analysis of a proper data structure that suits data formats and constraints ● Choice of adequate data structure based on the preliminary information about the data	*Competencies* ● Compare various data structures for a given problem, such as array, list, set, map, stack, queue, hash table, tree, and graph ● Compare the trade-offs of different representations of a matrix and common operations such as addition, subtraction, and multiplication ● Recognize data structures obtained after called script-based subroutines ● Evaluate how efficient data structure for the insert, remove, and access operations

Algorithms

A data scientist should recognize that the choice of algorithm will have an impact on the time and space required for a problem. A data scientist should be familiar with a range of algorithmic techniques in order to select the appropriate one in a given situation.

Computing Fundamentals: Algorithms	
Scope ● Problem solving through algorithmic, computational and statistical thinking. ● Algorithm design, implementation, and analysis. ● Comparison of various well-known computing algorithms' complexity, including machine learning (ML) and statistics techniques. ● Complexity of a given algorithm ● Factors that influence the algorithm complexity and performance ● Computational performance of certain algorithms based on providing different data sets.	*Competencies* ● Analyze the differences between iterative vs recursive-based algorithms ● Implement an efficient search algorithm to find a target with certain characteristics ● Provide the big Oh time and space for a given procedure ● Evaluate best, average, and worst-case behaviors of an algorithm. ● Apply an appropriate algorithmic approach to a given problem. ● Contrast which technique is more appropriate to use based on a given scenario

Software Engineering

Software engineering principles include design, implementation and testing of programs. A data scientist should understand design principles and their implications for issues such as modularization, reusability, and security.

Computing Fundamentals: Software Engineering	
Scope ● Software engineering principles, including design, implementation and testing of programs. ● Principles of object-oriented design such as encapsulation, inheritance and polymorphism to address concerns such as modularization, reusability and security; ● Principles of functional programming to maintain complex scaling applications and model/function composition ● Principles of compiled imperative programming for numeric computations and scientific computing. ● Probabilistic computing for testing and software lifecycle	*Competencies* ● Implement a small software project that uses a defined coding standard. ● Incorporate statistical models into the software lifecycle ● Evaluate results of a program by utilizing statistical significance testing ● Demonstrate how software interacts with various systems, including information management, embedded, process control, and communications systems ● Test a given piece of code by including security, unit testing, system testing, integration testing, and interface usability

Data Acquirement and Governance

There can be no analysis of data without the data itself. A data scientist must understand the source and quality of their data, as well as understand appropriate processes for acquiring and maintaining high quality data.

Data Acquirement and Governance	
Scope ● Shaping data and their relationships. ● Acquiring data from physical world and extracting data to a form suitable for analysis. ● Integrating heterogeneous data sources. ● Preprocessing and cleaning data for applications.	*Competencies* ● Construct and tune the data acquirement and governance process according to the requirements of applications, including the selection of data sources and data acquirement equipment, data preparation algorithms and steps. (Process Construction and Tuning) ● Define and write semantics rules for data acquirement and governance, including information extraction, data integration and data cleaning (Rules Definition) ● Develop scalable and efficient algorithms for data acquirement and governance according to the property of data and the requirements of applications, including data proper discovery, data acquirement, information extraction, data integration, data sampling, data reduction, data compression, data transformation and data cleaning algorithms (Algorithm Development) ● Describe and discover the static and dynamic properties of data, changing mechanisms of data and similarity between data. (Property Description and Discovery)

Data Management

A data scientist must understand the storage, maintenance, and retrieval of data.

Data Management	
Scope ● Storing and indexing (structured, semi-structured and unstructured) data ● Data models; query languages based on the data model ● Effective conceptual models and architectures for databases ● Data retrieval: queries, keywords, efficiency. ● Processing transactions in database management systems ● Scaling database management systems.	*Competencies* ● Design the logical and physical structure for effective data management according to data type, data model and application. ● Design index structure for efficient query processing and information retrieval. ● Describe the semantic requirements of data access in a declarative language or a keyword set. ● Tune and optimize the storage structure and query processing in data management systems for scalability and efficiency issues. ● Determine a strategy for transaction processing to balance efficiency, scalability and consistency of data management systems, especially for parallel and distributed environments. ● Design scalable and efficient algorithms for query processing, query optimization, transaction processing as well as information retrieval.

Data Privacy, Security, and Integrity

Data Privacy

This is intended to provide students with an understanding of data privacy and its related challenges. Students are expected to understand the tradeoffs of sharing and protecting sensitive information; and how domestic and international privacy rights impact a company's responsibility for collecting, storing, and handling data. **[xref: Professionalism: Privacy and Confidentiality]**

Data Privacy, Security, and Integrity: Privacy	
Scope ● Interdisciplinary tradeoffs of privacy and security. ● Individual rights and impact on needs of society. ● Technologies to safeguard data privacy. ● Relationships between individuals, organizations, and governmental privacy requirements.	*Competencies* ● Evaluate and understand the concept of privacy, including the societal definition of what constitutes personally private information and the tradeoffs between individual privacy and security. ● Summarize the tradeoff between the rights to privacy by the individual versus the needs of society. ● Evaluate common practices and technologies, and identifying the tools that reduce the risk of data breaches while safeguarding data privacy. ● Thoroughly comprehend how organizations with international engagement must consider variances in privacy laws, regulations, and standards across the jurisdictions in which they operate. This topic includes how laws and technology intersect in the context of the judicial structures that are present – international, national and local – as organizations safeguard information systems from cyberattacks.

Data Security

This focuses on the protection of data at rest, during processing, and in transit. This area requires the application of mathematical and analytical algorithms.

Data Privacy, Security, and Integrity: Security	
Scope ● Basic concepts in cryptography: Encryption/decryption, sender authentication, data integrity, non-repudiation; Attack classification (ciphertext-only, known plaintext, chosen plaintext, chosen ciphertext); Secret key (symmetric), cryptography and public-key (asymmetric) cryptography. ● Role of mathematical techniques for encryption. ● Role of symmetric (private key) ciphers for data security. ● Role of asymmetric (public-key) ciphers for data security. ● Cross-border privacy and data security laws. ● What are the data security laws and how do they impact. ●	*Competencies* ● Describe the purpose of cryptography and list ways it is used in data communications; and which cryptographic protocols, tools and techniques are appropriate for a given situation. Describe the following terms: cipher, cryptanalysis, cryptographic algorithm, and cryptology, and describe the two basic methods (ciphers) for transforming plaintext into ciphertext. Explain how public key infrastructure supports digital signing and encryption and discuss limitations/vulnerabilities. ● Exhibit a mathematical understanding of encryption algorithms, such as modular arithmetic, Fermat, Euler theorems, primitive roots, discrete log problem, primality testing, factoring large integers, elliptic curves, lattices and hard lattice problems, abstract algebra, finite fields, and information theory. ● Describe methods for data security, such as block ciphers and stream ciphers (pseudo-random permutations, pseudo-random generators), Feistel networks, Data Encryption Standard (DES), Advanced Encryption Standard (AES). ● Describe how mathematical concepts (such computational complexity) contribute to

	algorithms for data security. ● Explain requirements of the General Data Protection Regulation (GDPR); and Privacy Shield agreement between countries, such as the United States and the United Kingdom, allowing the transfer of personal data. ● Describe how certain laws [such as the following in the USA: Section 5 of the U.S. Federal Trade Commission, State data security laws, State data-breach notification laws, Health Insurance Portability Accountability Act (HIPAA), Gramm Leach Bliley Act (GLBA), and Information sharing through US-CERT, Cybersecurity Act of 2015] impact data security.

Data Integrity

Data integrity refers to the overall soundness, completeness, accuracy, and consistency of data.

Data Privacy, Security, and Integrity: Integrity	
Scope ● Approaches to the accuracy and consistency (validity) of data.	*Competencies* ● Explain the concepts of message authentication codes (HMAC, CBC-MAC); Digital signatures; Authenticated encryption, and Hash trees that provide data integrity.

Machine Learning

Machine learning refers to a broad set of algorithms and related concerns for discovering patterns in data, making new inferences based on data, and generally improving the performance of a software system without direct programming. These methods are critical for data science. Data scientists should understand the algorithms they apply, be able to implement them, if necessary, and make principled decisions about their use.

Machine Learning	
Scope • Broad categories of machine learning approaches (e.g., supervised and unsupervised) that make assumptions about the data available at learning time and the general types of inferences that can be made from that data. • Algorithms and tools (i.e., implementations of those algorithms) in each of the broad learning categories. • Machine Learning as a set of principled algorithms (e.g., optimization algorithms), rather than as a "bag of tricks." • Computational learning theory and what it tells us about the theoretical limitations of various approaches. • Notion of a hypothesis space of learning outcomes and its relationship to the expressive power of learned models. • Problems related to model expressivity as well as availability of data, and techniques for mitigating their effects. E.g., problem of overfitting and regularization techniques for mitigating effects of overfitting; curse of dimensionality and feature selection/weighting/reformulation techniques for mitigating effects. • Ways to evaluate performance,	*Competencies* • Compare and contrast broad classes of learning approaches, with a focus on inputs, outputs, and ranges of problem types to which they can be applied. • Select and apply a broad range of machine learning tools/implementations to real data. • Derive a (current) learning algorithm from first principles and/or justify a (current) learning algorithm from a mathematical, statistical, or information-theoretic perspective. • Express formally the representational power of models learned by an algorithm, and relate that to issues such as expressiveness and overfitting. • Exhibit knowledge of methods to mitigate the effects of overfitting and curse of dimensionality in the context of machine learning algorithms. • Provide an appropriate performance metric for evaluating machine learning algorithms/tools for a given problem. • Apply appropriate empirical evaluation methodology to assess the performance of a machine learning algorithm/tool for a problem. • Apply appropriate empirical evaluation methodology to

both in terms of specifying objectives (e.g., predictive accuracy, cost-sensitivity, size of learned model) and in techniques for measuring them. • Methodology for evaluating the model produced by a machine learning algorithm/tool for a single problem; methodology for empirically comparing algorithms against each other more generally. • Differences in interpretability of learned models. • Model drift over time. • Algorithmic and data bias, integrity of data, and professional responsibility for fielding learned models.	compare machine learning algorithms/tools to each other. • Implement machine learning programs from their algorithmic specifications. • Be aware of problems related to algorithmic and data bias, as well as privacy and integrity of data. • Consider and evaluate the possible effects -- both positive and negative -- of decisions arising from machine learning conclusions. • Compare differences in interpretability of learned models.

Data Mining

Data mining involves the application of machine learning and statistical techniques to extrapolate information from data.

Data Mining	
Scope ● Workflow of data mining and its relationship to data preparation and data management ● Data mining models for a variety of data models and applications ● Design and analysis of data mining algorithms for various data mining models	*Competencies* ● Design data mining models for specific data models according to applications (Model Design) ● Design a data mining system, including the system architecture, data process flow, and data storage structure (System Design) ● Develop efficient and scalable data mining algorithms for specific data models, data mining models as well as data management platforms (Algorithm Development) ● Evaluate the significance and usability of data mining results to ensure that they may be applied in real applications properly (Result Evaluation)

Big Data

The term 'Big Data' has been coined to describe systems that are truly large. These introduce problems of scale: how to store vast quantities of data, how to be certain the data is of high quality, how to process that in ways that are efficient and how to derive insights that prove useful. These matters are addressed below under the headings of problems of scale, complexity theory, sampling, and concurrency and parallelism.

Problems of Scale

Big Data: Problems of Scale	
Scope ● Approaches to storing vast quantities of data ● Ensuring clean, consistent and representative data ● Protecting and maintaining the data ● Retrieval issues ● Problems of computation and the efficiency of algorithms ● Specific techniques used in addressing the problems of scale	*Competencies* ● Explain the role of the storage hierarchy in dealing with Big Data ● Demonstrate how redundancy may be removed from a Big Data set ● Illustrate the role of *hashing* in dealing with Big Data ● Illustrate the role of *filtering* in dealing with Big Data

Complexity Theory

Big Data: Complexity Theory	
Scope ● The notion of computational complexity ● Limits to complexity ● Evaluation of the complexity of algorithms ● Selecting appropriate algorithms	*Competencies* ● Explain why mathematical analysis alone is not always sufficient in dealing with efficiency considerations ● Demonstrate how to evaluate the efficiency of an algorithm to be used in processing Big Data ● Select algorithms appropriate to a particular application involving Big Data, taking account of the problems of scale

Sampling and Filtering

Big Data: Sampling and Filtering	
Scope ● The role of sampling and filtering in the processing of Big Data ● Benefits of sampling / filtering ● Criteria to be used in guiding typical sample selection	*Competencies* ● Perform sample selection for a particular application involving Big Data ● List a variety of approaches to filtering, illustrating their use

Concurrency and Parallelism

Big Data: Concurrency and Parallelism	
Scope ● Concurrency and parallelism, and distributed systems ● Limitations of parallelism including the overheads ● Differing approaches to addressing concurrency ● Complexity of parallel / concurrent algorithms	*Competencies* ● Explain the limitations of concurrency / parallelism in dealing with problems of scale ● Identify the overheads associate with parallelism in particular algorithms

Analysis and Presentation

The human computer interface provides the means whereby users interact with computer systems. The quality of that interface significantly affects usability in all its forms and can encompass a vast range of technologies: animation, visualisation, simulation, speech, video, recognition (of faces, of hand writing, etc.) graphics. For the data scientist it is important to be aware of the range of options and possibilities, and to be able to deploy these as appropriate.

Analysis and Presentation	
Scope • Importance of effectively presenting data, models, and inferences to clients in oral, written, and graphical formats. • Visualization techniques for exploring data and making inferences, as well as for presenting information to clients. • Effective visualizations for different types of data, including time-varying data, spatial data, multivariate data, high-dimensional multivariate data, tree- or graph-structured data, and text. • Knowing the audience: the client or audience for a data science project is not, in general, another data scientist. • Human-Computer Interface considerations for clients of data science products.	*Competencies* • Explain data and inferences made from data in oral, written, and graphical formats. • Use standard APIs and tools to create visual displays of data, including graphs, charts, tables, and histograms. • Apply a variety of visualization techniques to different types of data. Make useful inferences / extract useful information from a dataset using those techniques. • Propose a suitable visualization design for a particular combination of data characteristics and application tasks. • Analyze the effectiveness of a given visualization for a particular task. • Describe issues related to scaling data visualization from small to large data sets. • Be aware that the client (for an interface or presentation) is often not a data scientist. • For an identified client, undertake and document an analysis of their needs.

Professionalism

In their technical activities, data scientists should behave in a responsible manner that brings credit to the profession. One aspect of this is being positive and proactive in seeking to bring benefit and doing so in a way that is responsible and ethical. Much of this is amplified in general terms in [1]. This section below serves to highlight a number of relevant issues of specific concern to the data scientist. A number of sub-areas are identified: continuing professional development, communication, teamwork, economic considerations, privacy and confidentiality, ethical issues, legal considerations, intellectual property, and on automation.

Continuing Professional Development (CPD)

The essence of a professional is being competent in certain aspects of data science. It is the responsibility of the professional to undertake only tasks for which they are competent. There are then implications for keeping up-to-date in a manner that is demonstrable to interested parties, e.g. employers.

Professionalism: Continuing Professional Development	
Scope ● The meaning of competency and being able to demonstrate competency ● Acquiring expertise / mastery or extending competency; the role of journals, conferences, courses, webinars ● Technological change and its impact on competency ● The role of professional societies in CPD and professional activity	*Competencies* ● Justify the importance to the professional data scientist of maintaining competence. ● Describe the steps that would typically have to be taken to extend competence or acquire mastery, explaining the advantages of the latter. ● Argue the importance of the role of professional societies in supporting career development.

Communication

There are various contexts in which the data scientist is required to undertake communication with very diverse audiences. That communication may be oral, written or electronic. There is the need to be able to engage in discussion about the role that data science can play, to communicate multiple aspects of the data science process with colleagues, to convey results that may lead to change or may provide insights. Being able to articulate the need for change and being sensitive to the consequences are important attributes. These activities may entail the ability to have a discussion about limitations in a certain context and to suggest a research topic.

The communication of the data scientist must be underpinned by an evidence-based approach to decision making. There is special significance to this in the context of machine learning and automation where the reasons for decisions may need to be clarified.

Professionalism: Communication	
Scope ● Different forms of communication – written, oral, electronic - and their effective use ● The technical literature relevant to data science ● Audiences relevant for communication involving the data scientist – including small groups, large groups, experts and non experts, younger groups, senior managers, machines – and the elements of effective communication in each case	*Competencies* ● Evaluate an aspect of the technical literature relevant to data science ● Produce a technical document for colleagues to use to guide technical development ● Design and present a case to senior managers outlining a major initiative stemming from a data science investigation

Teamwork

The data scientist will often act as a member of a team. This may entail being a team leader, or supporting the work of a team. It is important to understand the nature of the different roles as well as the typical dynamics of teams. So in terms of teamwork the data scientist needs to be able to collaborate not only with data scientists with different tool sets but, in general, with a diverse group of problem solvers.

Professionalism: Teamwork	
Scope ● Team selection, the need to complement abilities and skills of team members ● The dynamics of teams and team discipline ● Elements of effective team operation	*Competencies* ● Document and justify the considerations involved is selecting a team to undertake a specific data science investigation ● Recognise the qualities desirable in the team leader for a data science research investigation

Economic considerations

Data scientists need to be able to justify their own positions as well as the kind of activity in which they engage.

Professionalism: Economic considerations	
Scope ● The cost and value of high quality data sets, and the costs of maintenance ● Justifying in cost terms data science activity ● Estimating the cost of projects ● Promoting data science ● Automation	*Competencies* ● Predict the value of a particular data set to an organization, taking into account any requirement for maintenance ● Argue the case for the data that an organization should routinely gather ● Document the cost (in terms of resources generally) of collecting high quality data for a particular purpose ● Justify or otherwise the creation of a particular data science activity within an organization and quantify the cost ● Infer the value to an organization of undertaking a particular investigation or research project ● Document and quantify the resources needed to carry out a particular investigation in house and compare this with outsourcing the activity ● Evaluate and justify the costs associated with the automation of a particular activity

Privacy and Confidentiality [xref: Data Privacy, Security, Integrity]

It is possible to gain access to data in a multitude of ways, by accessing databases, using surveys or questionnaires, taking account of conditions of access to some resource, and even with developments such as the Internet of Things, specialized sensors, video capture and surveillance systems. Although gaining access to all kinds of information is important, this must be done legally and in such a way that the information is accurate and the rights of individuals, as well as organizations and other groups, are protected.

Professionalism: Privacy and Confidentiality	
Scope ● Freedom of information ● Data protection regulations including GDPR – see [5] ● Privacy legislation ● Maintaining the confidentiality of data ● Threats to privacy and confidentiality ● The international dimension	*Competencies* ● Describe technical mechanisms for maintaining the confidentiality of data ● Compare the privacy legislation in two specific countries, and indicate the problems arising from the differences ● Recognize the privacy and confidentiality issues arising from the use of video and face recognition software

Ethical issues

Ethical issues are of vital importance for all involved in computing and information activities. Such issues are captured extensively in [1]. Underpinning these is the view that a professional should undertake only tasks for which they are competent, and even then should carry out such tasks in a way that reflects good practice in its many forms. Maintaining or extending competence is essential. A heightened awareness of legal and ethical issues must underpin the work of the data scientist.

Teaching students to consider the ethical issues associated with their decisions is a very important starting point, enabling them to recognize themselves as "independent, ethical agents."

Professionalism: Ethical Issues	
Scope ● Confidentiality issues associated with data and its use ● The General Data Protection Regulation (GDPR) regulation – see [5] ● The need for data to be truly representative ● Bias in data and in algorithms; mechanisms for checking	*Competencies* ● Demonstrate techniques for establishing lack of bias in a set of data or in algorithms ● Create a technical paper on an aspect of data science for colleagues ● Reflect on a network of professionals in the data science area and outline the advantages to be gained by joining such a network

Legal Considerations

Computer crime has continued to increase both in volume and its severity over recent years. This has brought disruption, even chaos, to many organizations. The threat of computer crime cannot be ignored and steps need to be taken to counter the possibility of severe disruption. The law has adjusted to counter these trends but this is an ongoing area of change and adjustment.

Professionalism: Legal Considerations	
Scope ● Computer crime – examples of most relevance to data science ● Cyber security ● Crime prevention ● Mechanisms for detecting criminal activity, including the need for diverse approaches ● Recovery mechanisms and maintaining 100% operation ● Laws to counter computer crime	*Competencies* ● Illustrate and evaluate a range of mechanisms for detecting a stated form of criminal activity ● Justify the desirability of having multiple diverse approaches to countering threats

Intellectual Property (IP)

Intellectual Property rights such as copyright, patents, designs, trademarks and moral rights, exist to protect the creators or owners of creations of the human mind; moral rights include the right to be named as a creator of IP, and the right to avoid derogatory treatment of creations. For the data scientist the items to be protected, in possibly different ways, include software, designs (including GUIs), data sets, moral rights and reputation. Trade secrets may also be relevant.

Professionalism: Intellectual Property	
Scope ● patents, copyrights, trademarks, trade secrets, moral rights ● what data science related IP can and cannot be protected, and what kinds of protection are available ● regulations related to IP, IP ownership, the territorial nature of IP rights including the effects of international agreements (e.g. the European Directive on trade secrets) and the issue of IP rights being time limited ● which IP rights vest automatically and which require registration, including overview of the processes involved in acquiring registered IP rights ● the possibility of infringing the rights of others and validly utilizing protected IP	*Competencies* ● Describe what kinds of IP are relevant to the data scientist and why ● Argue the difference between patents, copyrights, designs and trademarks and illustrate their use in the context of data science ● Describe the role of trade secrets in relation to data science ● Illustrate the processes involved in registering IP rights ● Describe and explain the issues relating to IP ownership and moral rights ● Evaluate the risks involved in using protected IP and how they may be validly overcome

Change Management

One possible outcome of a data science investigation is that strategic change is needed in an organization. The change suggested may be minimalistic at one extreme or transformational at the other. The data scientist needs to be alert to the range of possibilities and, perhaps by engaging other experts, be in a position to offer advice and guidance about how to move forward with such change, to advise on the consequences and to outline and quantify the resources that will be required to deliver on the change.

Professionalism: Change Management	
Scope ● The need for strategic change, the role of simulation ● Structural change, transformational change ● Strategies for delivering effective change including top-down and bottom up approaches ● Resource needs associated with change ● People issues associated with change, managing resistance to change, the role of human resources and communication ● Monitoring the effectiveness of change ● The role of automation	*Competencies* ● Justify with evidence the need for strategic change within an organisation and recognise the nature of the change required (e.g. personnel change, structural change, transformational change) ● Provide a set of feasible approaches about how to deal with transformational change in a given situation, to quantify the resources needed and to highlight benefits ● Outline a range of strategies for managing resistance to change ● Identify teamwork, leadership and personnel issues associated with change including gaining support from management, maintaining operation while implementing change and addressing loss of employment ● Recognise the issues associated with change brought about by automation (including ethical issues), with possible need for back-up

On Automation

Automation often creates concerns about loss of employment opportunities and, in general terms, about machines behaving unreasonably; explanations from machines about their behavior may be sought. Related issues are the subject of [3] and [6]. Automation can occur in critical situations where serious loss may be possible, and then typically there is an expectation that machines will operate to a code of ethics in sympathy with that of humans.

Professionalism: On Automation	
Scope ● Automation, its benefits and its justification ● The particular concerns of automation in critical situations ● Transparency and accountability in algorithms	*Competencies* ● Analyze the impact on design of the requirement to provide insights into decisions made autonomously by machines ● Argue the benefits of automation in particular situations

References

[1] The ACM Code of Ethics and Professional Conduct, published by ACM on 17th July 2018. See acm.org

[2] When computers decide: European Recommendations on Machine-Learned Automated Decision Making, published by ACM, 2018. See europe.acm.org

[3] ACM US Public Policy Council and ACM Europe Policy Council, "Statement on Algorithmic Transparency and Accountability," 2017.

[4] Directive (EU) 2016/943 on protection of undisclosed know-how business information (trade secrets) against their unlawful acquisition, use and disclosure. See eur-lex.europa.eu June 2016.

[5] The EU General Data Protection Regulation, see www.eugdpr.org. Approved by EU on 14th April 2016 with an implementation date of 25th May 2018.

[6] Simson Garfinkel, Jeanna Mathews, Stuart S. Shapiro, Jonathan M. Smith, Towards Algorithmic Transparency and Accountability, Communications of the ACM, September 2017,vol. 60, no. 9, page 5.

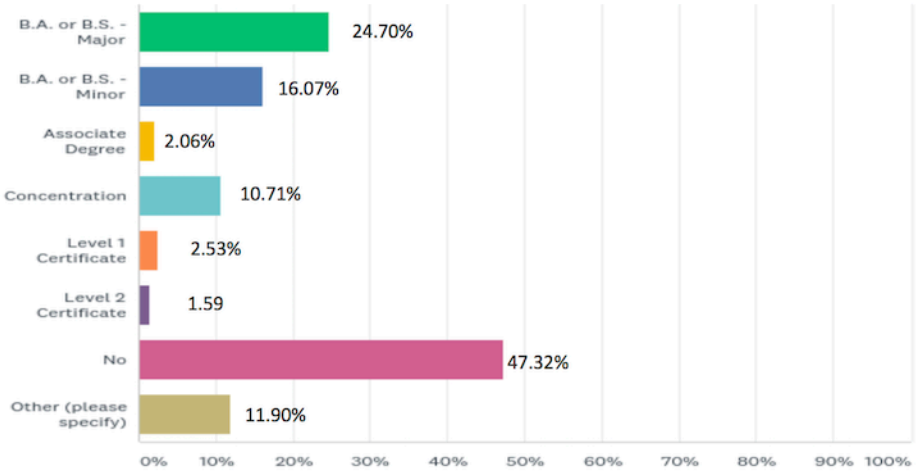
Appendix B: A Summary of Survey Responses

Here we include a subset of responses to the Academic and Industry surveys.

B.1 Academic Survey

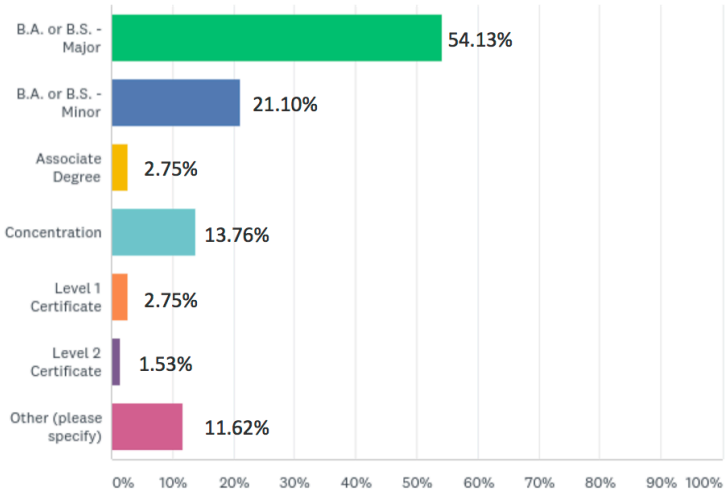
Q1: Does your institution offer an undergraduate program in Data Science or Analytics? (Check all that apply.)

Answered: 672 Skipped: 0



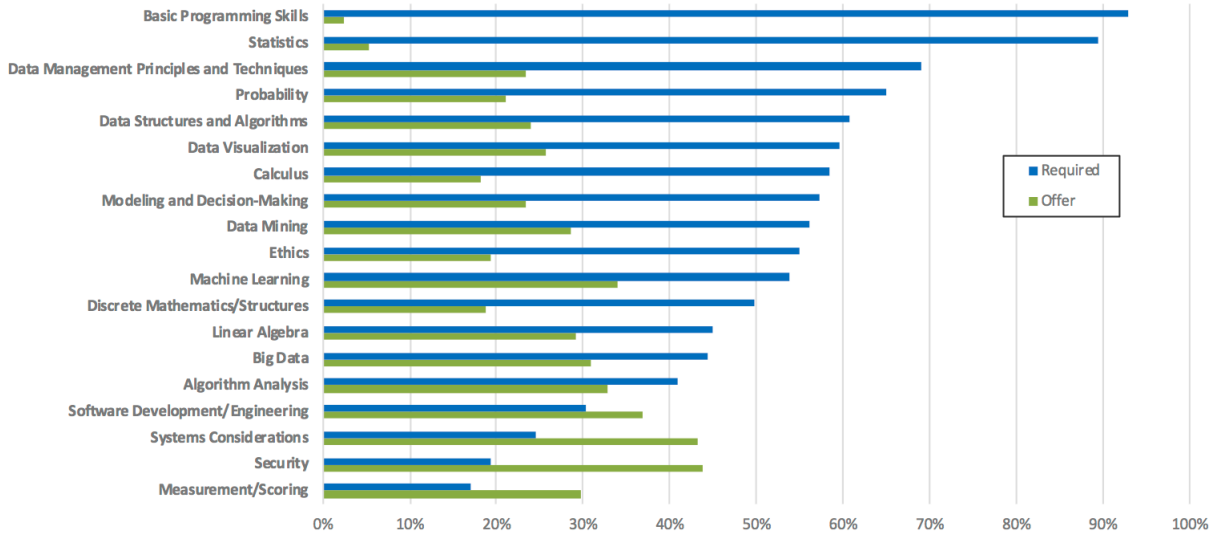
Q2: Please select one program for which you will answer a set of curricular questions.

Answered: 327 Skipped: 345



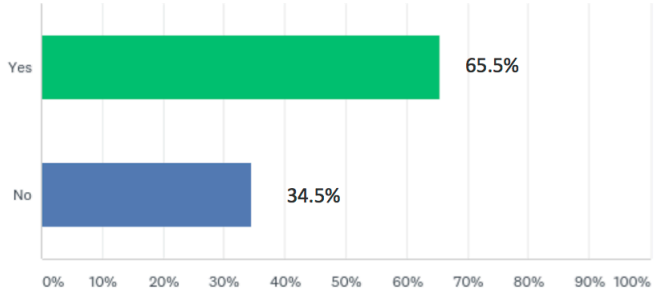
Q4: Does your program offer/require content from these areas?

Answered: 172 Skipped: 500



Q6: Does your program have a "data science in context" requirement? (i.e., a requirement to apply data science methodology to some area of application)

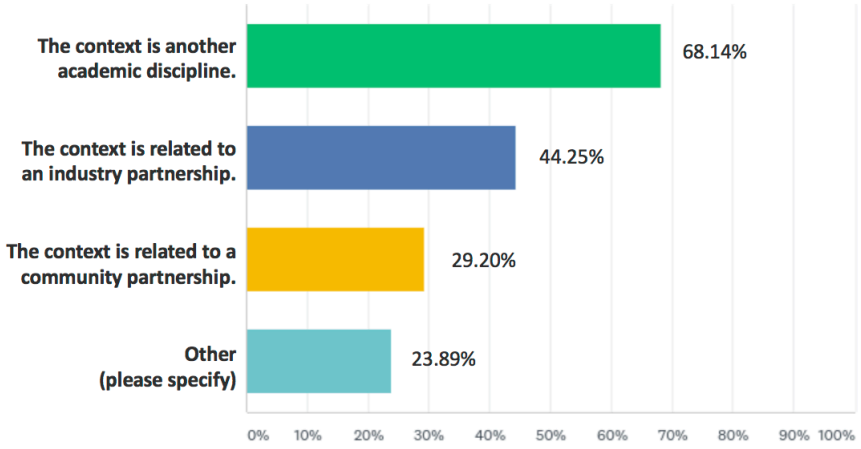
Answered: 171 Skipped: 501



Q7: Check all that apply

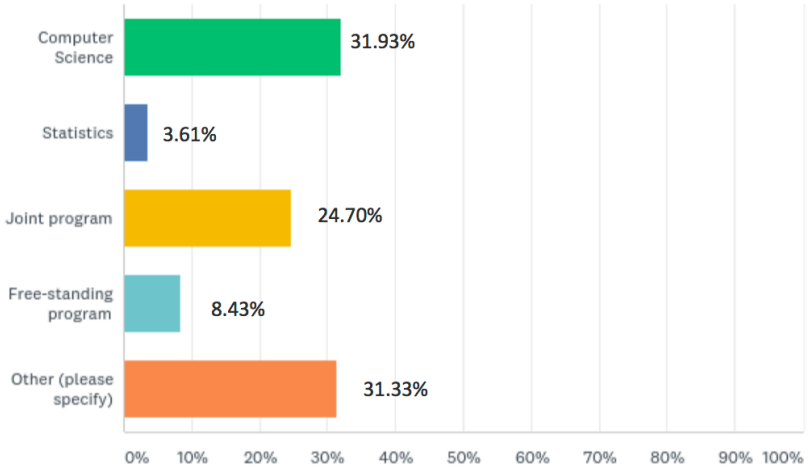
For: Does your program have a "data science in context" requirement?

Answered: 113 Skipped: 559



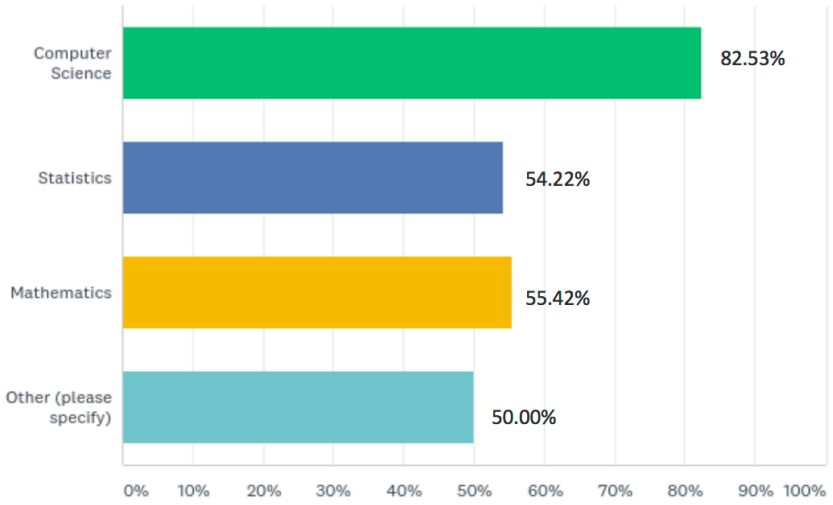
Q8: What is the academic home of your Data Science/Analytics program?

Answered: 166 Skipped: 506



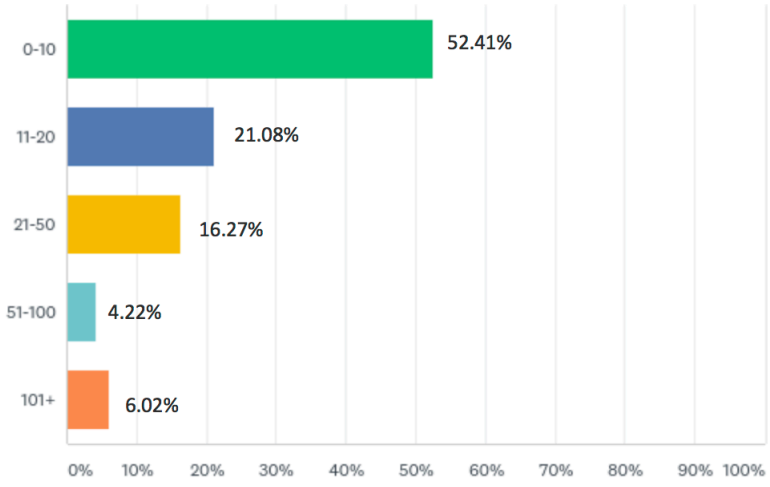
Q9: What academic units/departments contribute to your Data Science/Analytics program?(Check all that apply.)

Answered: 166 Skipped: 506



Q10: How many students graduate with this degree annually?

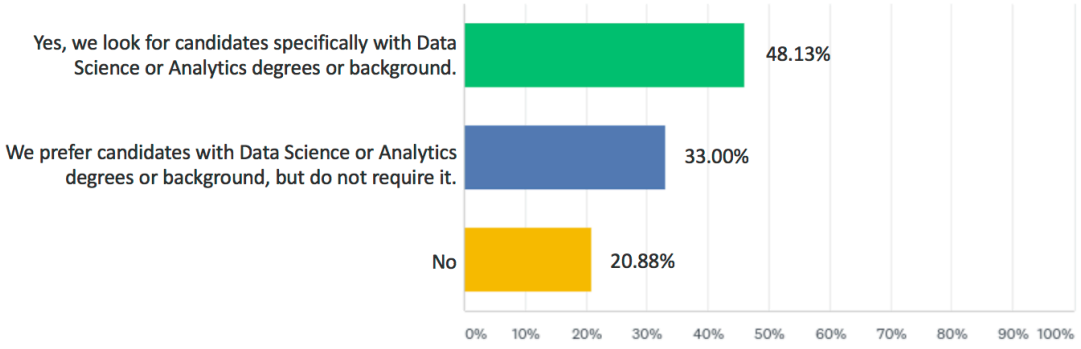
Answered: 166 Skipped: 506



B.1 Industry Survey

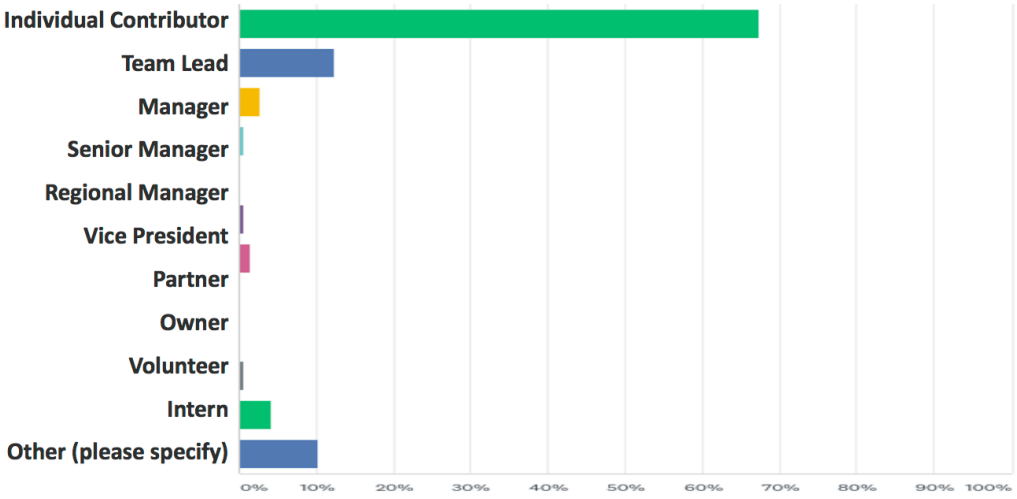
Q1: Do you look for job candidates (specifically new graduates out of undergraduate programs) with Data Science background?

Answered: 297 Skipped: 0



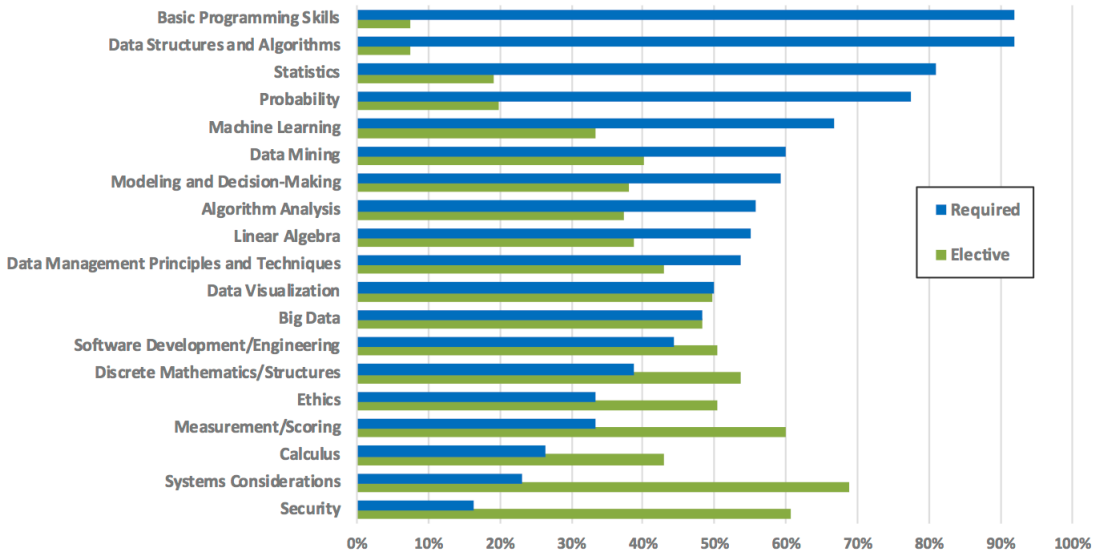
Q2: Provide the title of the job for which you require/prefer Data Science/Analytics degrees or background. The next few questions on the survey will pertain to that position.

Answered: 147 Skipped: 150



Q3: For the job title you provided, to what extent do you require experience in the following areas?

Answered: 147 Skipped: 150



Q4: In the list of COMPUTING-based data science areas above, what did we miss?
Open-Ended Response

AI, knowledge representation, text analytics, machine learning in perception, database (including NoSQL)
Analytical redundancy Binary standard - ramification
Basic applied domain understanding
Causal inference
COLAS
Data Driven Control Systems
Data analysis and interpretation; decision support for executives.
Data integration
Data warehouse management principles.
Design Thinking: the ability to find insights that matter into huge datasets
Graph theory
Heavy hands-on experiences. Exposure to very difficult problems.
Implementation. Life cycle management of the data science/machine learning algorithms.
Ontology engineering
Pattern recognition, Knowledge generation, General Intelligence, data gathering
Python SAS R
Simulation
SQL coding
Strong communication skills.
Study design and Interpretation of findings (These two are related).
Textual analytics
Too many data science education programs focus on mathematics, but not enough on actual computer programming.

Q5: Do you expect a job candidate to have experience applying data science in context? (i.e., experience applying data science methodology to some area of application)

Answered: 143 Skipped: 154