Strengthening Data Science Education Through Collaboration

October 1-3, 2015

Report on a Workshop on Data Science Education

Funded by the National Science Foundation, Award #: DOE 1545135

Boots Cassel, Villanova University (cassel@acm.org)

Heikki Topi, Bentley University (htopi@bentley.edu)

Executive Summary

This document provides a report on the background, process, findings, and recommendations of a Fall 2015 workshop on Data Science Education. This workshop was funded by the National Science Foundation and organized by the ACM Education Board and Council on October 1-3, 2015 in Arlington, VA.

Almost 200 colleges and universities offer degrees in data science at the time of writing, and this number is appears to be increasing rapidly. A wide variety of schools and departments provide these degrees, including business, computational science, computer science, mathematics, and statistics. Data science is "a set of *fundamental principles that guide* [emphasis added] the extraction of knowledge from data… Data science involves principles, processes and techniques for understanding phenomena via the (automated) analysis of data." [3] Perspectives on data science do, however, vary according to the viewer and to the particular kinds of problems to be solved.

One of the key challenges preventing organizations and societies from gaining full benefits that data science could potentially offer is the shortage of professionals with a sufficient educational background. Despite the large number of educational programs that have emerged to address this shortage, our collective understanding of the quality and future direction of these programs and the best structures and practices for them is limited. This workshop was designed to start a conversation to address these concerns and develop a deeper integrated understanding of the best ways to offer data science education, ultimately leading to a better prepared workforce.

The workshop discussions addressed questions related to the definition of data science, the core knowledge and skill areas of data science practice, the relationship of data science with application domains, implications and potential consequences of data science practice, developing a workforce for data science, and the actions needed to move data science education forward.

In a summary form, key thematic findings included the following:

- Data science can either be seen as an integrated combination of contributing disciplines or a separate, still emerging discipline with its own content and expected competencies. Many participants supported the view that data science is gradually emerging as a discipline that is different from the academic disciplines that contribute to it. Data science should be freed from the contributing disciplines so that it can gain its own identity. For this to be possible, data science will need to develop its own independent theory and methods.

- Discussion regarding data science is challenging because various participating communities use the same terms with different meanings. It is essential that we not end up trapped by using language with a well-defined prior meaning to describe a new phenomenon — the old language might become a constraint for progress.
- Data science offers profoundly important new opportunities for transforming human activity at various levels. Data is foundational, pervasive, tacit, and often collected without a specific intent. Achieving the benefits is not trivial. Data science is a process, including all aspects of gathering, cleaning, organizing, analyzing, interpreting, and visualizing the facts represented by the raw data. All the aspects are critical to meaningful results, and every step presents opportunities for lost or misused data leading to erroneous or even deliberately misleading results. Collaboration between academic organizations and partners representing organizational practice is essential, and the current climate is favorable for supporting this collaboration.
- Understanding the ethical implications and potential consequences of data science processes is critically important. Data science practitioners need the skills to evaluate the implications of their work from the ethical perspective.
- The workshop did not come to an agreement regarding the need for a core. The lack of a clear understanding of what data science is contributes to the conflicting views about the core. Without a more mature, shared understanding of the identity of the discipline, it will be hard to define a core and a broader curriculum plan. Participants expressed concern about the possibility that a premature attempt to define a core might inhibit development of the field. Flexibility related to topics emphasized and care in specifying a core emerged as important requirements for any follow-up actions.
- Broad topic areas that emerged as central in all discussions regarding essential topics and competency areas in data science include machine learning, statistical inference and modeling, data and database management, data integrity, privacy, and security, and the domain of activity for which data is important.
- Given the emerging nature of data science as a discipline, its development is affected by traditional university structures that often do not support the integrative approach that data science requires. In a specific university context, it is possible that competition for resources and organizational politics can significantly impede the speed of progress. Top university management leadership is essential for enabling the integrative structures that are needed.

The workshop recommends that:

1. A broadly interdisciplinary task force be formed and sufficiently funded to develop curriculum guidance for degree programs in data science. This task force should consist of representatives from at least ACM (representing computer science), ASA (statistics), AAAS (cross-section of sciences), AIS (business domain), INFORMS (management science and operations research), and to be identified organizations representing humanities and social sciences as domains.
2. Infrastructure and culture of sharing of materials and experiences (including negative ones) among the departments and schools that offer data science programs be supported and encouraged. We should strive to form a knowledge hub across several faculties, domains of knowledge and industry partners, a hub that offers visibility and connections between many existing platforms. Forms of collaboration might include web-based tools, conference(s), or a journal focused on data science education.

Table of Contents

Introduction

This document provides a report on the background, process, findings, and recommendations of a Fall 2015 workshop on Data Science Education. This workshop was funded by the National Science Foundation and organized by the ACM Education Board and Council on October 1-3, 2015 in Arlington, VA.

Background

According to one source, degrees in Data Science appear in the offerings of almost 200 colleges and universities in the United States. More than 100 additional degrees are available in other countries (http://datascience.community/colleges), and the number of them keeps increasing. A wide variety of schools and departments provide these degrees, including business, computational science, computer science, mathematics, and statistics[1]. Some are listed as interdisciplinary. The general definitions of data science are broad and open to many interpretations.

The Wikipedia definition of data science gives us a broad view of the field:

> "In general terms, **Data Science** is the extraction of knowledge from data. It employs techniques and theories drawn from many fields within the broad areas of computer science, information theory and information technology, and mathematics and statistics – including signal processing, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression and high performance computing.. Methods that scale to Big Data are of particular interest in data science, although the discipline is not generally considered to be restricted to such data. The development of machine learning, a branch of artificial intelligence used to uncover patterns in data from which predictive models can be developed, has enhanced the growth and importance of data science."

Provost and Fawcett [7] define Data Science as phenomenon at a higher level of abstraction:

[1] The names of the departments and disciplines are listed alphabetically here and later in the document.

> Data Science is "a set of *fundamental principles that guide* [emphasis added] the extraction of knowledge from data… Data science involves principles, processes and techniques for understanding phenomena via the (automated) analysis of data."

Perspectives on data science vary according to the viewer and to the particular kinds of problems to be solved. As Dhar states, "the term "science" implies knowledge gained through systematic study."[3] The data may be extremely large, perhaps making it prohibitive to move the data around for various kinds of processing. The data may be of many kinds: binary, text, images, audio and video, representations of locations in space and time, etc. The data may vary in terms of its sensitivity and need of high degree of protection, both for accuracy and for confidentiality. Data formats may be incompatible with other data sets that need to be merged. For many in science, data science is part of the Fourth Paradigm of Science [8], in which discovery happens through interactions with the data—organizing, structuring, cleansing, and then analyzing.

Data science is often considered in relation to other, related academic fields and domains of practice. For example, workshop participants representing statistics suggested that statistics could be thought of as the science of learning from data in general whereas data science leverages computational capacities more explicitly. Some observers see a close connection between data science and big data analytics, the latter of which is often defined based on characteristics of data or sufficiency of techniques and technologies available to domain experts at a particular point in time.

Programs offered in business schools and departments often focus on a variant of data science called business analytics. This subject overlaps data science in that it also seeks to make effective use of large quantities of information, but the focus is on "the study of data through statistical and operations analysis, the formation of predictive models, application of optimization techniques and the communication of these results to customers, business partners and colleague executives." (http://www.stern.nyu.edu/programs-admissions/global-degrees/business-analytics/program-overview/what-business-analytics)

Given the broad range of disciplines that are contributing to the practice of data science, it is not surprising that a rich variety of academic disciplines are serving an important role in data science education and that the organizational models for building and offering data science programs vary significantly. Departments and independent centers are competing for power and resources in this area that is currently receiving keen interest from a variety of stakeholders. Data science programs are rooted in different departments, engage different mixes of instructor actors, draw upon dissimilar curricula, and

target varied student groups. All of these factors inhibit the development of a robust body of knowledge and curriculum for data science.

Motivation and Key Goals

The importance of data science as one of the key drivers of advances in science, business, and government has been recognized broadly in recent public discussion. One of the key challenges preventing organizations and societies from gaining full benefits that data science could potentially offer is the shortage of professionals with a sufficient educational background. A large number of educational programs have emerged to address this shortage, but our collective understanding of the quality and future direction of these programs and the best structures and practices for these programs is limited. This workshop was designed to start a conversation to address these concerns and to develop a deeper integrated understanding of the best ways to offer data science education, ultimately leading to a better prepared workforce. Workshop results will be made available on the Ensemble Computing Education Portal and other similar platforms for broad access within the computing community and also shared through other disciplinary portals to reach the wider community.

The goals of the meeting included the following:

- Creating an opportunity for key experts in data science education to discuss fundamental issues in this area without the disruptions caused by internal political considerations or immediate resource competition
- Allowing primary academic and professional societies that have a stake in data science education to share their goals and experiences and explore opportunities for collaboration
- Developing tentative recommendations regarding the following questions:
    - What are the most useful practices for allowing early stage data science programs to share information so that they can learn from each other (including learning from failures)?
    - Would it be helpful to develop a curriculum recommendation for undergraduate programs and/or master's programs in data science?
    - Would it be beneficial to consider accreditation criteria for undergraduate programs in data science?

Participants

The participants of the workshop included the following experts representing a wide variety of academic fields and professional and academic societies with a strong interest in data science education:

Anderson, Paul (University of Charleston)
Beck, David (University of Washington)
Carson, Cathryn (UC Berkeley)
Cassel, Boots (Villanova University, workshop co-chair)
Culler, David (UC Berkeley)
Getoor, Lise (UC Santa Cruz)
Goul, Michael (Arizona State University)
Hancock, Stacey (UC Irvine)
Horton, Nick (Amherst College)
Kraska, Tim (Brown University)
LaLonde, Donna (American Statistical Association)
Lee, Dongwon (Penn State, NSF)
Mentzer, Kevin (Bryant University)
Peckham, Joan (University of Rhode Island)
Phillips, Jeff (University of Utah)
Plale, Beth (Indiana University)
Posner, Michael (Villanova University)
Prey, Jane (ACM representative)
Schedlbauer, Martin (Northeastern University)
Sun, Heshan (Clemson University, AIS representative)
Sykora, Martin (IEEE-CS representative, NexJ)
Topi, Heikki (Bentley University, workshop co-chair)
Washington, Anne (George Mason University)
Wilson, Jill (Northwestern University, INFORMS representative)
Wixom, Barb (MIT Sloan School of Management)

Initial Questions to Participants

When the participants were invited to the workshop, they were asked to start to consider the following questions:

- Just what is meant by the term data science? What other terms are related, with overlapping or conflicting meanings? For example, we note that there is a lot in common between business analytics and data science, but there are differences also. Are there identifiable common themes and easily described places where the two are different?

- How much of data science is about the theory and concepts related to handling data, independent of the application domain? Can we talk about data integrity, privacy, security, storage and manipulation, sharing, distributed processing, etc. without caring about the purpose the data serves?

- What are the specific theories and concepts that provide needed tools and techniques for those faced with mountains of data and needing to know what it can tell them? Whether it is genetics, astronomy, survey-based social science data, or customer behavior, what approaches will yield the knowledge needed to make decisions, design products, or set new directions for study?

- What are the general and specific needs for those preparing to work as data scientists to be able to understand and evaluate the implications and potential consequences of their work?

- Does data science have a sufficiently mature conceptual foundation and educational practices to allow the community to consider the development of a curriculum recommendation? Would accreditation criteria and processes be relevant and necessary?

The goal was also for the meeting to produce specific recommendations, including whether or not it would be beneficial to have an undergraduate curriculum recommendation for data science, and whether or not it is time to consider accreditation criteria.

Pre-workshop Work by the Participants

Before the workshop, the participants were asked to respond to the following questions:

> *What do you believe to be the most important competences of an early career data scientist? In your opinion, which ones of these competences can be developed by university education and which ones are best to be developed in an employment context?*

Current university degree programs in data science are offered at the master's level and at the undergraduate level. In your experience, does one level appear more appropriate than the other? If so, what level? What leads you to that conclusion?

Many computing disciplines have developed a body of knowledge either as part of a curriculum development project (such as the ACM/IEEE and ACM/AIS initiatives for Computer Engineering, Computer Science, Information Systems, Information Technology and Software Engineering) or separately (such as SWEBOK). If you were considering a similar effort for data science, what would be your top three to five choices to be included as top level categories (Knowledge Areas)? If possible, please describe each area with a couple of sentences. If you want to include more than five Knowledge Areas, it is perfectly fine, but please indicate the top five.

Do we know enough about desirable data science program characteristics to develop a model curriculum and/or accreditation criteria? Would such resources aid in the maturation process for the field? Please explain your reasoning.

Please list up to three "must read" references to your own or others' works related to data science education or data science in general. For each, please include a 1-2 sentence annotation explaining why others "must read" it. If you can send us an electronic copy of these readings, we will make them available on a password-protected webpage for workshop participants.

Appendix A includes the responses the participants provided to these questions.

Structure of the Workshop and Key Topics For Each Session

The sessions of the workshop and their core questions were as follows:

Session 1: What is Data Science?

Core questions:

- We do not have to agree and we do not expect the group to agree regarding the fundamental identity of data science. It is, however, important for us as a group to set the stage by understanding and articulating the richness of perspectives on what data science is. The purpose of this session is to explore how many different ideas of data science do we, as a group, have. We would also like to find out whether or not we can bring the perspectives together into a reasonably small set.

- Where do the divisions show up? Is it based on the initial academic discipline (computer science, information science, information systems, statistics, etc.), the application domain (business, sciences, humanities, etc.) or the environment in which data science is practiced (academic, research lab, business, etc.)

Introductory presentations during this session were given by Cathryn Carson, Lise Getoor, Michael Goul, and Michael Posner.

Session 2: What are the core knowledge and skill areas of Data Science practice?

Core questions:

- What are the absolutely necessary knowledge and skills areas forming the core of data science practice? (From the proposal: What are the specific theories and concepts that provide needed tools and techniques for those faced with mountains of data and needing to know what it can tell them? Whether it is genetics, or astronomy or social science data, or survey results, or customer behavior, what approaches will yield the knowledge needed to make decisions, design products, or set new directions for study?)
- What are the special areas that apply to your department/group within your own university? Do you believe there are forms of data science outside your own unit's definition?

Session 3: Relationship of Data Science with application domains

Core questions:

- How much of data science is about the theory and concepts related to handling data, independent of the application domain? How much should students of data science learn about data integrity, privacy, security, storage and manipulation, sharing, distributed processing, machine learning, statistical methods, algorithms and data structure, etc. without caring what purpose the data serves? Is it meaningful to have a data science degree without an application domain? How do we determine the balance between various disciplinary perspectives?
- Do the answers vary depending on context, environment? Are there things we can agree on?

Introductory presentations during this session were given by Paul Anderson, David Beck, Nick Horton, and Barbara Wixom.

Session 4: Implications and potential consequences of Data Science practice

Core questions:

- Can we teach aspiring data scientists to understand implications and potential consequences of their work? Should ethics of data science be included in data science programs?

Session 5: Developing a workforce for Data Science

Core questions:

- What are the most essential steps that need to be taken to develop a competent workforce in data science?
- What types of programs in data science do we need?
- Who should offer these programs?
- What roles can professional societies play?

Session 6: What actions are needed to move Data Science education forward?

- Is there enough agreement about at least some needs to suggest a curriculum recommendation effort?
- Is it time to consider an accreditation effort?
- Who could lead efforts in curriculum design?
- Who could accredit programs in data science?

Appendix B includes the detailed workshop schedule. Workshop presentation materials are available at http://bit.ly/DSEW2015.

Thematic Findings

Defining Data Science

The workshop participants made a conscious decision not to spend a significant amount of time *defining* data science. If we had attempted to create a definition, we would not have been able to discuss any other issues during the entire workshop. The integration of general scientific principles, computer science, information science, mathematics, statistics, and subject matter expertise creates an intersection that has

as many definitions as there are academics attempting to define it.

The conversations during the workshop did, however, produce interesting observations regarding what data science is, including the following:

- There are at least three essential definitional questions related to data science:

    1. What is data science?
    2. What is unique about data science?
    3. Is data science a new discipline?

- The potential definitions vary from very simple (such as "Data science is about making decisions with data" or "The fundamental question of data science is: How do we do things smarter with data?") to highly complex questions regarding the ways multiple scientific disciplines are integrated to form a new area of study.

- Data science includes a broad range of elements in addition to actual analysis of data, starting from identification of sources of data and ending with communication of analysis results and changes of behavior or practices based on the analysis. Data science is a process, including all aspects of gathering, cleaning, organizing, analyzing, interpreting, and visualizing the facts represented by the raw data. All the aspects are critical to meaningful results, and every step presents opportunities for lost or misused data leading to erroneous or even deliberately misleading results.

- We can look at data science as a separate field, with its own content and expected competencies; alternatively, we can look at it as a marriage of existing disciplines, drawing from each to make something distinct, but still closely connected to its component parts.

- Many participants supported the view that defining data science requires that it be extracted from all the areas from which it is emerging. Data science is gradually emerging as a discipline that is different from the academic disciplines that contribute to it, and it has to be freed from the contributing disciplines for it to get its own identity.

Data Science and Analytics

One of the significant definitional questions is the relationship between data science and other, related areas of study. A particularly important set of related areas consists of different forms of analytics (such

as business analytics, legal analytics, music analytics, etc.). No specific solutions emerged from the workshop to provide criteria for making the distinction (or determining whether or not a distinction is needed). Some participants expressed a strong opinion that, for example, business analytics and data analytics are different from data science and serve different purposes. Others equated data science and analytics stating, for example, that "Data science has been around for a long time under many different titles, such as data analytics." However, it was not clear what criteria could be used to draw a distinction between the two or to demonstrate that they are essentially the same.

Challenges with Human Language/Terminology

Given the different disciplinary backgrounds, it is difficult to discuss data science at a high level of specificity because the same foundational terms are used with different meanings. The participants agreed that learning how to communicate about data science and its results across traditional disciplinary boundaries is essential, but also very difficult. The choice of words is important when we describe the movement forward, particularly because use of discipline-specific language may have political connotations. Using language that indicates a strict selection of one option is problematic and is likely to scare away potential participants and contributors.

It is, however, essential that we not end up trapped by using language with a well-defined prior meaning to describe new phenomena—the old language might become a constraint for progress. The term "interdisciplinary" was cited as an example. One piece of advice given during the workshop was: "The strategy that seems to work sometimes is to be just really frank and direct in the language. Say things like *funky* and *wonky*. Put that in the report."

Reasons Underlying the Importance of Data Science

The workshop participants, all of whom were chosen because of their active involvement in data science education, all naturally shared a strong belief in the importance of data science. The motivation for this was not, however, based on narrow disciplinary considerations. Instead, it was justified by broad societal reasons. We live now in a world where everything an individual or an organization does creates a digital footprint, including the actions of communicating, purchasing, and any physical or virtual movement. Data is foundational, pervasive, tacit, and often created without specific intent. This creates great opportunities for improving the overall quality of a variety of human activities through data, but achieving the benefits is not trivial. To be beneficial, the data needs to be "wrangled like crazy," and it can be easily demonstrated that many potential uses are not beneficial. Therefore, those responsible for

using data to change ways in which humans behave have to be technically highly qualified and must understand the implications and potential consequences of data.

Motivation, Excitement, and Momentum

The workshop identified a variety of reasons why it is exciting to be working in the context of data science at this point in time. The opportunities created by ubiquitous data lead to a significant demand for data scientists in a variety of fields of human activity; this, in turn, requires that more data scientists be educated. Some organizations are even stating that data science could "change our entire existence."

The excitement expressed about working in data science appeared to be fueled, at least in part, by the very newness of the kinds of work done. The connection between the hard facts in the data and a sense of enhanced knowledge derived from the data motivates and energizes this community.

In certain contexts, the excitement may lead to unjustified hype. For example, in popular press, data science is seen as magic: "It's going to make your business profitable. It's going to cure cancer. It's going to turn mud into gold." The group also identified the methodological tension that emerges when two cultures of reaching conclusions from data meet or crash into each other, one that assumes a specific stochastic data model and another one that is based on algorithmic models [2]. This tension leads at times to uncertainty regarding the validity of the approaches used by the other culture.

Overall, the participants of the workshop expressed their own excitement about the opportunity to be working in a field that is currently emerging with its own identity. It is highly motivating to be working on something new that offers numerous opportunities to affect a variety of fields of human activity positively. Mixed with the enthusiasm was a strong sense of urgency regarding the need to address the educational needs of various stakeholder groups (business, government units at various levels, not-for-profit organizations, scientific disciplines, etc.) related to data science. Many stakeholders are developing their own training in this area to address short-term needs. The academic community has the special responsibility to provide education that develops a lasting foundation and allows graduates to adapt to changing circumstances.

Inherently Interdisciplinary Nature

A very common theme throughout the entire session was the interdisciplinary nature of data science. The participants agreed that data science is broadly interdisciplinary, despite some misgivings about the adequacy of that term. There are at least two primary reasons for this: First, as discussed frequently

above, data science brings together multiple methods and disciplines. Second, data science is an applied discipline that integrates the methods with the processes and practices of a domain of human activity, such as a scientific discipline, business, or government. There was some discussion about whether data science can exist as a field of study independent of a domain of application. Data science was seen as fundamentally integrative in nature. Some participants ventured as far as to say that the "magic of the workshop group" was the diversity of the fields represented.

Interestingly, several participants viewed data science as a framework or platform for bringing experts from multiple disciplinary perspectives together. This was characterized with observations such as:

- Data science is a really good term for bringing people together. We should care less about what the specific content is and be less concerned about the specific content than the *actual process*.
- Data science is a platform for bringing interesting people together, as experienced at UC Berkeley not only in education but also in research.
- Data science leads to a diversity of perspectives and interesting people coming together across the widest possible range of domains.

Some participants suggested that we should, ultimately, go beyond interdisciplinary. We need an approach that does not care about boundaries of traditional disciplines in the problem solving process -- the focus should be on formulating problems and solving them ("no boundary research, no boundary thinking")[5]. There is a need for new interdisciplinary theory and methods, recognizing the fact that different disciplines are doing similar things but using different terminology or language. Ultimately, something new is emerging — a data science program cannot just be an integrated collection of existing courses. We cannot simply "tie a ribbon around some things that exist" — doing that will not lead to an optimal outcome.

There are existing examples of areas where multiple disciplines have come together to form something different. One of them is HCI (Human-Computer Interaction). HCI has certainly a computing component, but there is also an art component and a psychology component. The ACM SIGCHI has fewer than half its membership in the computing field. HCI programs do not mix existing computing courses with existing art and psychology courses. Instead, they have defined their integrated field over time, and they have an active education mission.

As will be discussed later, there was also a strong consensus regarding the fact that the interdisciplinary nature of data science causes challenges that need to be actively managed. For example, it is essential to

include the domain disciplines, but finding a way to accomplish this requires forms of collaboration that are new to most of us.

Curriculum

Programs in data science are emerging with a wide variety of goals and differing views of what data science is. Guidance in defining programs must account for these differences and respect the cultural differences as well as variations in institutional infrastructure that drive choices. There is room and even a growing desire for help in defining the core elements that every program must have. Beyond that, any curriculum effort must include options to address specific needs.

Should We Define a Core?

Defining a core for data science will not be easy. There is a widespread agreement that a core would include elements of statistics and elements of computer science. As discussed above, data science is inherently interdisciplinary. The process of defining a core must find ways to describe topics, or perhaps to define competencies, without extensive references to existing disciplinary language.

The participants' perspectives regarding the importance of the core varied significantly from statements as strong as "There has to be a core" to "I'm going to invite you to keep resisting that impulse to come up with a core and say 'We know what this should be'" (because the future is going to bring new things). Others believe that the core should be defined in a way that is different from the way in which curriculum cores have been defined earlier (although it was not necessarily clear how the new way would differ from the old one). It is clear that the diversity of existing programs and the differences that stem from the perspectives of the different disciplines contributing to the definition present challenges to reaching an agreement on the nature of the field and the essential elements of any definition of a core.

The lack of a clear understanding of what data science is contributes to the conflicting views about a core. Without a more mature, shared understanding of the identity of the discipline, it will be hard to define a core and a broader curriculum plan. Participants expressed concern about the possibility that a premature attempt to define a core might inhibit development of the field. Flexibility related to topics emphasized and care in specifying a core emerged as important requirements.

It is also essential that we keep in mind the multitude of competencies that graduates of data science are expected to have. The expectations also vary significantly depending on the specific focus of their desired or anticipated career path. The workshop participants emphasized the fact that broad definitions of data

science and data scientist cover a wide range of professional tasks and performance expectations. Exactly the same curriculum would not be able to support the preparation needed for building systems, developing novel analytic methods, exploring models to describe trends and relationships, preparing data for analytical purposes, etc.

Finally, when we discuss the core, it is essential that we make it clear what elements of the curriculum we are referring to: Core disciplines, core competencies, core topics or maybe core knowledge areas?

The Broader Topic List for Data Science

Machine learning always appears in the discussion regarding required topics. However, machine learning itself is a big topic that spans computer science and statistics, requiring capabilities from both disciplines. Curriculum development will require deconstructing machine learning to extract the essential elements. Reduced to its fundamentals, machine learning provides and develops techniques to classify or cluster large data sets using efficient algorithms for both solving optimization problems and representing the solutions for inference [1]. Statistical evaluation of the results determines how much confidence the results deserve. What, then, does a typical data scientist need to know about machine learning? Perhaps it is important to know what the models represent and how to choose one over another for a given purpose. It still leaves the question of how much of the computational model or the statistics behind the results must the data scientist understand and how much to be able to develop independently.

Overall, topics and competencies related to statistics and statistical inference were identified as central elements of data science. Statistics provides an essential set of competencies related to the processes of reaching conclusions from data, particularly in understanding the effects of randomness and quantifying uncertainty. Solid foundation in statistics will also decrease the danger of misunderstandings related to the types of questions that a particular analytical approach can address [6].

Data management also appears consistently in lists of core competencies for a data scientist. That, too, is a large topic that multiple existing disciplines (at least computer science, information systems, and information technology) compete to own. Components include data collection, with associated ethical issues; data integration; data "munging" or "wrangling"; organizing data, perhaps in a database; creating schema; writing queries; interpreting query responses. Visualization and other approaches to communicating conclusions from the data are also essential. There is importance not only in the individual steps in gathering and managing data, but in the sequence of steps that leads reliably to a valid conclusion.

Curriculum areas include data integrity, privacy, and security in addition to data collection, manipulation, and presentation discussed above. These also rely on algorithms, data structures and statistical methods. Other topics that appear in data science programs include natural language processing, data mining, artificial intelligence, and visualization. Some elements appear as tools to be used, not to be created. Others feature strongly in tasks for the students to accomplish. As will be discussed later in this report, the participants also emphasized the importance of incorporating the development of ethical principles and an understanding of the implications of data science methods and practices in data science education.

In every discussion of data science curriculum, the domain of the data appears as a dominant factor. The question arises as to the feasibility of assembling topics and competencies independent of a specific domain. Abstract ideas of data management, independent of any problem domain, do not appear to develop the needed instincts and skills. A related question is the transferability of data science competencies developed in one domain to use in another. Is that often possible? Are there circumstances that support transferability or that inhibit it? Does it matter what the domains are? Is there a fundamental underlying similarity between finding evidence of a planet in astronomy data and finding evidence of fraud in financial data?

Repeatedly, the workshop attendees asserted that all data scientists should be grounded in a domain, not in the methodology alone. Contextualization and project-based learning supports the integration of methodology and domain.

Comprehensive Topic List

In addition to the most frequently identified focus areas specified above, the workshop identified many others. The following integrated list includes the competencies and/or knowledge areas that were featured in discussions of data science curriculum at the workshop:

- Knowledge of algorithms, programming, software engineering
- Machine learning
- Statistics, experimental design, exploratory data analysis, ability to make valid inferences, understanding causality
- Communication with very diverse audiences, oral and written, including the ability to communicate multiple aspects of the data science process. The ability to have a discussion about the research topic and the limitations
- Teamwork skills, ability to collaborate not only with data scientists with different toolsets but in

general, with a diverse group of problem solvers [4]

- Evidence-based decision making or problem-solving
- Ability to dive deep and solve problems
- Decision support
- Creativity, being creative within the constraints of given instructions
- Ability to identify categories and patterns
- Data shaping, data wrangling, data munging—getting the data from its source and putting it into a form suitable for analysis. Understanding the origins of data, data preparation, data monitoring, data curation. Management and preparation of heterogeneous types and high volumes of data.
- Visualization
- Data privacy, security, integrity
- Teaching students to consider the ethical consequences of their decisions is a very important starting point. Enabling students to recognize themselves as "independent, ethical agents." Providing students with tools for ethical analysis.
- Classification
- Philosophy of information
- Managing the purpose -> data -> insight -> action -> value process
- Understanding of the unavoidability of model decay and acting on it when it is observed
- Analytics, sometimes written with perspective of specific domains, and
- Data mining simulation.

The inherent interdisciplinary nature of data science raises issues about the order in which aspects appear and how tightly they are coupled. For example, some participants suggested that students studying bioinformatics become very competent in that domain, but are not prepared to transfer their data science competencies to other domains. For them, data science and biology may be so tightly integrated that they cannot see the bigger picture.

Platforms for Sharing

Communication is an essential component of the development and evolution of data science programs. As programs emerge in very different contexts, innovations and experiences could easily be lost and have to be rediscovered. The need for communication includes sharing what works so it does not have to be reinvented; and also sharing what has been shown not to work.

There are several existing platforms for sharing materials. These provide access to teaching resources, curriculum ideas, and stories about experiments. Making connections among these and other sites can help bring resources to larger communities. These sites include the following:

- teachingdatascience.org
- computingportal.org
- teradatauniversitynetwork.com

Sharing experiences regarding options that did not work is more difficult than simply sharing resources. There are risks associated with admitting to failure. Establishing a safe way to share experiences that did not work out as planned would be a valuable contribution. Defining metrics for success and failure makes this possible. However, we do need a culture that recognizes value of lessons learned, including and particularly lessons learned from challenges and difficulties. This will be further discussed in our final recommendations.

Changing Status Quo within Universities

Workshop participants expressed a shared concern regarding the traditional university structures that do not support the integrative approach that data science requires. These structures have the potential to inhibit the collaboration and integration that is needed for data science to thrive. Overall, universities have not been able to figure out how to deal with interdisciplinary programs and initiatives within the typical department structure. New structures are currently being established—such as centers and institutes and also joint programs with rotating chairs—that have been designed for these purposes. Structures are significantly affected by institutional constraints—to use the "let a thousand flowers bloom" metaphor that emerged often during the workshop, we need to recognize the environmental differences affecting the growth of our rose.

It is clear that that both in research and in teaching, new opportunities created by access to large amounts of data also required bringing together multiple specialties in ways that did not exist before. The change needed is more substantial than simply identifying multiple disciplinary players and dividing responsibilities between and among them. We need to reorganize and reconfigure the disciplines themselves in a way that matches the new reality. This is a challenge for universities, funding agencies, and professional societies. We might, indeed, be running the risk of "locking ourselves into structures that are coming from the history." This is natural, given that many mechanisms for resource allocations are linked to existing structural arrangements, and those who control these resources have a strong reason to defend the status quo.

For example, in the case of data science, for many universities a crucial question appears to be whether data science efforts should be led by computer science or statistics. Given the important role that many other academic disciplines play in data science, the question is not a simple either/or question; much more sophisticated sharing mechanisms are needed.

Finally, the workshop identified a typical boundary within universities that is relevant also for this conversation: this boundary exists between those faculty and staff members who advocate for keeping things the way they have always been and those who are motivated by the desire to advance research and education, regardless of the changes it requires. This is often a clearer boundary than that existing between different disciplines.

Power, Politics, and Resources

The discussions at the workshop revealed a number of specific issues related to organizational decision making and resource allocations that potentially have a major impact in the universities' ability to develop and offer data science programs. In defining a field of study that is deeply dependent on existing fields but at the same time clearly interdisciplinary in nature, there is a danger of seeing the connections between the participating (and necessary) partners as competitive, rather than complementary. If resources are scarce, an "us vs. them" mentality may limit the quality of the integrative potential. The size of the core in a field is necessarily limited; what must be given up if something new is added? At the same time, universities that have the required resources and organizational structures to offer high-quality programs in data science and its derivatives (currently highly sought out areas of study) have the potential to increase their overall resource base and serve their external stakeholders better.

In practice, the speed and ease at which a new program in data science can be created depends significantly on the role of the academic administration (provost, deans) in the process. If the motivation to create a new major or a new degree program comes from the top, the process can happen quite quickly, particularly if it is supported with strategic injections of resources to create incentives for all critical players to be involved. Otherwise, it can be difficult and time consuming to create the right conditions for collaboration.

Ideally, the drive to create programs should come from top down and bottom up at the same time. Equally important is that the drivers for developing data science programs do not only include expected surpluses from the new programs. Instead, it is essential for universities to recognize the need to incorporate concepts of data science into university curricula so that the academic community has the necessary intellectual resources to support other academic areas in dealing with the major challenges and

opportunities with data that all of them have.

Creating interdisciplinary structures includes significant political challenges, particularly related to the ownership of particular topics and areas of study. One of the most politically contentious issues related to data science is the role that each of the methods areas (CS, mathematics, statistics etc.) should play. Based on shared experiences, it appears that it is important to manage these relationships explicitly. For example, a data science program with a long history at University of Charleston is a joint program between mathematics and computer science. All decisions need to be approved by both departments.

Given the essential role that application domains play in data science programs, the program(s) should not be automatically hosted by a single department (which tends to be either CS or statistics). Universities should recognize that data science can either be placed centrally within a university (balancing the needs and requirements of statistics and computer science) or in a decentralized way with the domains. They can also consider new organizational structures, such as data science institutes, that might play a significant role in offering the programs and support research simultaneously. For example, University of Washington has an eScience Institute (http://escience.washington.edu/). It is essential that a new program has a strong institutional unit or a powerful individual as a champion; otherwise, there is a risk that it will be discontinued in the face of first signs of adversity, such as resource constraints.

Role of Ethics and the Ability to Understand Implications of Data

One of the general themes that emerged a number of times during the conversations was the role of ethics and the competences related to understanding the implications of new uses of data in data science education. Overall, there was a strong general sense that understanding ethical implications should be an important part of data science education.

This is an area where the concerns of academics and industry practitioners are fully aligned. For example, one of the participants described the results of a study on the views of chief data officers and other executives in similar roles. In this context, many of the top issues were related to ethics and implications. For example, the executives had recognized that customers perceive data science methods and techniques too pervasive and even "creepy." There were also significant concerns regarding the risks involved in automated, algorithmic decision making that is beyond the visibility of those ultimately responsible for the decisions.

Some of the concerns expressed at the workshop go beyond the "pretty tame scenarios of profit and loss." These ethical consequences and challenges include, for example, those related to the use of data in the

context of automated warfare. Ultimately, data scientists have to accept accountability for the decisions made as an end result of their work and therefore, their preparation must include elements that develop the competencies related to identifying ethical issues and making decisions that carefully consider ethical implications.

It is also important to recognize the connection between security and ethics: in many situations, ensuring an ethically sound outcome requires a high level of security.

Industry Collaboration

Data science is an applied field that benefits from connection with the professional practice of the application areas. Corporate partners can provide invaluable information regarding the desired and achieved educational outcomes. Some of the major questions in this area are as follows:

- True collaboration needs structural mechanisms that systematically support it, such as advisory boards. These mechanisms help academics understand industry needs and improve industry understanding of our goals and constraints. Academics do not expect or want them to direct efforts or make demands, but it is essential that two-way conversations take place in an environment of trust.

- Industry partners are not happy with what they are currently experiencing. At least some of them view the quality of the educational services that universities provide to be substandard. The ones happiest with universities are those that have created programs in close collaboration with the universities. Industry partners are also frustrated by the silo structure of universities. For example, they might have multiple groups simultaneously initiating conversations with academic units regarding the same issues.

- The collaboration opportunities are not only limited to large corporations. Indeed, small and medium size companies account for a major share of the data science work. MIT has done a good job of engaging such companies. Start-ups are good partners because they are often more willing to share data than established corporations governed by strict policies.

- It would be helpful if programs were able to engage industry better in capstone projects and internships. Intellectual property rights often cause challenges in these efforts.

One mechanism through which individual programs can get a broader view of industry needs than that emerging from conversations with their own advisory boards are reports from research centers (when

publicly available). For example, MIT's CISR serves its member companies with applied research. Recently, it conducted a large-scale study on executive perspectives on value of data. This study identified two broad categories articulating ways to create value: by using data to support primary assets of the company or by monetizing data directly. The study also discovered that corporations are concerned about similar types of issues as academic organizations, including the centralization/decentralization decision. Some workshop participants considered best practices of industry collaboration sufficiently valuable to indicate the need for a separate study and report.

Follow-up and Recommendations

The workshop participants represented a broad cross section of the areas concerned with data science, both from the standpoint of the subject areas and from the industry and academic perspectives of the questions. Through a voting process, the workshop participants indicated nearly unanimous support for an effort to develop curriculum guidance for data science. This led naturally to questions about who should be involved in that effort. There have been partnerships among multiple organizations with a shared interest in a specific curriculum area in the past, and that model has been successful (such as the collaboration between ACM and IEEE-CS and ACM and AIS on computing curricula). The data science effort will be likely to require different partnerships. The need for representation of the methodology areas (computer science and statistics) and also representation of a broad collection of domain areas will make the collaboration more complex than previous ones.

Thus, the first recommendation of the workshop is that a broadly interdisciplinary task force be formed to develop curriculum guidance for degree programs in data science. This task force should consist of representatives from at least ACM (representing computer science), ASA (statistics), AAAS (cross-section of sciences), AIS (business domain), INFORMS (management science and operations research), and to be identified organizations representing humanities and social sciences as domains. It is essential that the work of this task force should be sufficiently funded. Given the complexity of interdisciplinary and cross-organizational work, the process needs funding from outside the participating organizations. The workshop encourages National Science Foundation to enable this effort financially.

It is important to recognize that some participants of the workshop warned against codifying a specific recommendation too early. Interest in experimentation and exploring new directions remains high, and it is essential that specifying curriculum guidance should not become an obstacle for the development of new, innovative content or pedagogical models. It is also important that the guidance should not be framed normative in a way that would set unnecessary boundaries for the field's development or the

emergence of an entirely new discipline from the evolving integration of the component disciplines. This requires an agile process that accepts the idea of frequent changes and builds structures that are extendible and easily maintainable and changeable.

The second recommendation of the workshop is the development of both infrastructure and culture of sharing of materials and experiences among the departments and schools that offer data science programs. We should strive to form a knowledge hub across several faculties, domains of knowledge and industry partners. As discussed earlier, there are currently several platforms that offer the technical resources for sharing materials, but this is not sufficient. We need better visibility and links between the platforms; different communities have a legitimate need for their own environments for sharing, but it would be important to have a mechanism through which members of different communities would easily learn about content available within the others. Also, the community as a whole needs increased awareness of the resources that are available and encouragement and incentives for sharing materials, models, pedagogical approaches, and experiences, both positive and negative. It would be particularly important to find ways to enable and encourage sharing of failures so that others could learn from them. It is possible that web-based environments are not sufficient for building a culture of sharing; in addition, an interdisciplinary conference, symposium, or workshop might be beneficial, as could a journal focused on data science education.

Acknowledgements

Boots Cassel and Heikki Topi as co-chairs of the workshop and ACM Education Council and Board as organizing units want to express their deep gratitude to the participants of the workshop who were willing to provide their highly valuable time, effort, and insights to make this process possible. The organizers also want to thank the National Science Foundation, ACM, Bentley University, and Villanova University for the financial support without which the workshop would not have been possible. This material is based upon work supported by the National Science Foundation under Award # DOE 1545135. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2014.
[2] L. Breiman, “Statistical Modeling: The Two Cultures,” *Statistical Science*, vol. 16, no. 3, pp. 199–

231, 2001.
[3] V. Dhar, "Data Science and Prediction," *Commun. ACM*, vol. 56, no. 12, pp. 64–73, 2013.
[4] L. Hong and S.E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *PNAS*, vol. 101, no. 46, pp. 16385-16389.
[5] X. Huang et al. "No-boundary thinking in bioinformatics research," *BioData Mining*, vol. 6, no. 19, 2013. Available at http://www.biodatamining.org/content/6/1/19 , accessed July 26, 2016.
[6] J.T. Leek and R.D. Peng, "What is the question?" *Science*, vol. 347, no. 6228, pp. 1314-1315.
[7] F. Provost and T. Fawcett, *Data Science for Business*. O'Reilly Media, Inc., 2013.
[8] S. Tansley and K. M. Tolle, *The Fourth Paradigm*. Microsoft Press, 2009.

Appendix A: Pre-workshop questions and integrated participant responses

This section includes unedited responses given by the participants to the pre-workshop questions. The material has been utilized in the body of the report above.

Question 1: What do you believe to be the most important competences of an early career data scientist? In your opinion, which ones of these competences can be developed by university education and which ones are best to be developed in an employment context?

1. An ability to formulate "messy" problems and determine an analytical approach, if appropriate 2. The ability to identify the data needed to approach the problem and have the skills to harness that data (to include data management and cleaning) 3. The ability to glean insights from large data sets through the use of descriptive and predictive analytics 4. The ability to those insights into actionable items, for example, through prescriptive analytics 5. The ability to convince an organization to implement those actions. This requires understanding of organizational change and strong communication skills I believe that we can teach all of these to some extent in an academic context—especially 1, 3, and 4—but all will be refined on the job.
Mastering big data requires a set of skills spanning a variety disciplines from distributed systems over statistics to machine learning, a deep understanding of a complex ecosystem of tools and platforms, as well as communication skills to explain advanced analytics. I believe, the most important skill for a Data Scientist is to understand how these techniques work together and to be able to speak the language of the different disciplines.
The Data Science Education Working Group at the University of Washington has identified four core competencies, which are represented as the core classes required for the transcriptable Data Science options on the UW campus. They are Statistics, Machine Learning, Data Management (Databases), and Visualization. All of these can and should be introduced in the university education system. In addition, I have come to believe that there is an additional component that must be introduced before any of the above: Software Engineering. Too often, we focus on teaching programming languages or algorithms, but the rules and guidelines for writing software are ignored. This leads to poorly written one-off code that can neither be generalized nor reproduced. In my view, then, it is appropriate to begin instruction with Software Engineering before proceeding to the rest of the core competences. As for development in an employment context, all of the competencies can be refined in employment context, but need to be in place before leaving university. Another quick note… There is room for professional and continuing education to play a role in post-hoc re-education. I think this is an important area to seed the data science community with experienced domain people who want to go in new directions.

To me a data scientist needs to have natural curiosity and an affinity to solving puzzles. They need to be competent in using computational tools, including programming languages such as R and Excel, statistics, and facility with packages. Programming languages and programming concepts can be taught at the University-level along with statistical and machine learning techniques. Curiosity cannot be taught but can be fostered through experiential learning and project-based courses that have applicability to a student's domain.
Core competencies are statistical thinking, computer programming, visualization, communication, and domain knowledge. The best case scenario is probably iterative that is competencies are addressed in formal educational setting and augmented and reinforced in the employment context. The employment context will likely be more important for developing the required communication and domain knowledge competencies.
I define a data scientist broadly as someone with the technical knowledge and tool skills for extracting useful insights from the variety of data generated in today's digital economy. The process to extract insights often involves the use of statistics or mathematical models that seek to explain data element interactions with the fewest parameters. Key is the ability to provide structure to a problem so as to work around the limitations of the models / questions / data to reach actionable conclusions. It seems to me that, once out of the classroom, the fit between business requirements, the models and the available data becomes strained. Understanding the business questions to be answered and how to structure business problems in a way that will provide reliable insight is a key requirement. Ability to communicate solutions and more importantly caveats effectively. Perhaps an alignment around corporate performance management may in time make this process easier, but the communication boundaries between the technical execution channels and the management absorption channels can be broad and awkward to cross. Quite often analytics are defined with a domain of interest such as Marketing, Finance, Human Resources or Manufacturing. Defining clear metrics within each lines of business provides definition, guidance, expectations and delivery clarity as each line of business leader will need to understand why each metric is important and what its value to the company is. The technical articulation and implementation of that metric may be very complicated but it must satisfy the overarching need specified. Value increases further when dealing with areas like Supply Chain Management that cross lines of business, ensuring crisp alignment of metrics within and across each provider. Note that a certain degree of overall Business Intelligence maturity is required and expected before embarking on a metric journey, but it can provide the basis for more future facing questions that data scientists are often called upon to answer. I believe that the communication aspects could be taught in an academic context reasonably effectively – though I do not know that they are always being covered that way. I think some programs over focus on technical skills, leaving the communication aspect as more of a check-box exercise. Likely the best route would be to encapsulate best practices from successful consulting engagements and use that content as a basis for a course curriculum. An involved community where practitioners are called upon to contribute, update and refine these best practices over time will cull any deficient items from the knowledge domain and increase awareness on the need for new skills that need extra attention either in the educational setting, the workforce setting or the consulting practice setting. This mechanism should be applied similarly to dealing with ethics and processing of sensitive or personal data where the community keeps the practitioners in check.
The most important skills are to be able to think critically about several aspects of analysis and

scalability. Specifically, what is and what is not possible with tools from machine learning and data mining. What are critical bottlenecks in scalability from both an algorithm design perspective as well as on the implementation side of data management. And how to visualize data effectively without obscuring or biasing the truth. In each area students should be taught the core fundamental techniques in detail, and explain the challenges in extending them. Students should also learn modern terminology and notation so that they can pick up, read, comprehend, and assess new material as it becomes developed in the future.

I strongly believe that specific software tools should not be required, as they will become stale. Although python and R are both becoming great "catch-alls".

Not everyone needs to have every skill listed below. A broad education in many aspects, that drills down in depth into at least one.

- Data analysis (including but not limited to computational, mathematical, statistical, and engineering approaches)
- Data management
- Data archival
- Data access
- Ethical/Social aspects of data

I believe that organizations will ultimately take a federated approach to data management; some infrastructure and data management capabilities will be done at scale for the entire organization, but there will be tenets of data science within domains, and the nature of data/analysis for one domain will require specific skills and experience not required of the entire organization. I suspect universities will respond with evolving programs that will only develop general competencies targeting the organization-level capabilities, that some programs will develop general competencies but then go beyond in a specialization domain, and there will be some programs that target (or become well known for) a particular domain with specialized technologies – admitting students who already have prior general competencies. The first archetype (GENERAL) will address ETLT, data warehousing, math/stat, business performance management, predictive modeling, machine learning model applications, streaming data methods, temporal data methods, some prescriptive analytics that enable deployment within business processes and applications, etc. The next archetype (GENERAL + SPECIALTY) includes all competencies from the first archetype plus a deep dive in one or more of the specialty areas within a domain. The third archetype will be more like a teaching hospital – admitting already competent data scientists who wish to specialize in a domain, the methods and techniques suited to the specialty, plus the hands-on engagement only possible from being embedded in the domain (TEACHING HOSPITAL METAPHOR). All three archetypes will be present in university education.

Others at the workshop will fill out the intellectual foundations and technical skills. Let me highlight the **other** things that stand out when observing data scientists at work.

* Practical experience applying data science approaches in some particular domain. Which domain may not matter much. (To be provocative, what do we make of the fact that industry seems no less happy right now – maybe more happy? – to hire PhDs in astrophysics or

neuroscience than CS undergrads.) What's important here is partly the hands-on experience tackling the messiness, entanglement, and back-and-forth refinement process of a domain-area problem, and the extirpation of either a "one tool to rule them all" or a "twiddle the knob" mentality that comes from having experienced that there are trickinesses and alternatives to any seemingly obvious solution.

* Basic collaboration skills, like learning how to respect and play off of others with different expertise, how to divide and conquer and come back together, and how to troubleshoot as a team. Some appreciation of workplace dynamics and organizational context.

* A bottom-line understanding that no data problem is ever tidy, that there are always things that are yet to be understood. Also some kind of operationalizable awareness that the manner of data collection makes a difference for everything that comes after it, that there is no such thing as "pure" data that just corresponds to what's simply out there in the world.

* Awareness that there are huge and tricky practical and ethical problems around privacy, anonymization, other questions of use, and the ways in which poorly done analyses carried out with the authority of "data science" can powerfully damage people in the real world. Also an at least theoretical appreciation that the right solution may have to be worked out in an organizational, social, or political setting (i.e., it's not just what the technical expert might think is right).

Caveat: I'm assuming a "data scientist" is someone who **applies** data science approaches to real data in order to get actionable knowledge about things in the world, whether that's in business/industry settings or in university research or wherever. That's not an uncontested definition – my colleagues doing foundational work in CS and Statistics sometimes say they're data scientists, too – but it seems to be where this workshop is headed.

There are three elements important for data science – computing skills to assemble and manage data, statistical thinking to make valid inferences, and subject matter expertise. Analytic tools are crucial, which originated in the statistical domain but have been adopted more recently within the computing domain. These can be developed by university education, although not all of them need to be.

Management communication

Critical thinking, problem solving, scientific method

Data acquisition and management

Statistics – descriptive

Statistics – discovery

Artificial intelligence

Domain area

Process management, change management

Software engineering

Query languages: SQL/No SQL

Reporting and visualization

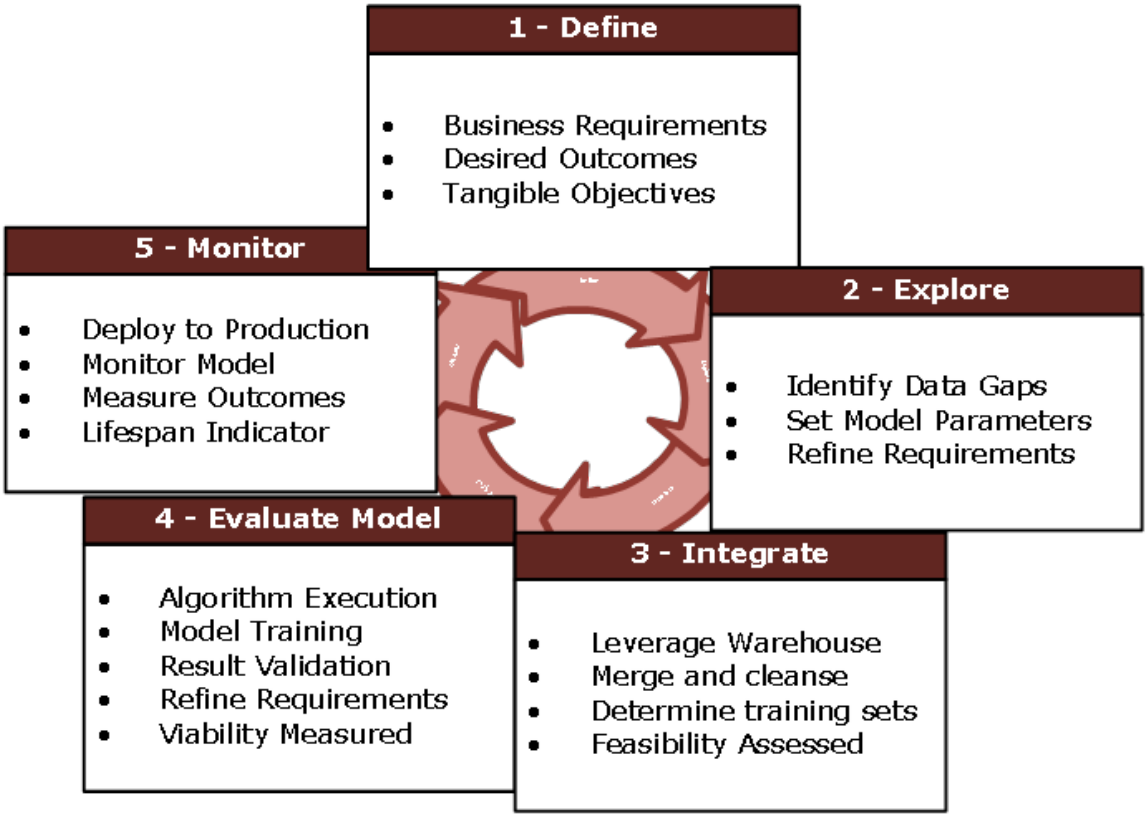
Data ethics, security, and controls

I would hope that a university could help begin developing all of the above skills. Then, a data scientist would continue to build and hone skills in an employment context.

Question 2: Current university degree programs in data science are offered at the master's level and at the undergraduate level. In your experience, does one level appear more appropriate than the other? If so, what level? What leads you to that conclusion?

It depends on what level of sophistication you want workers to have. I certainly think that undergraduate students can learn basic descriptive and predictive analytics and can learn to become consumers of analytics techniques at many levels. However predictive and prescriptive analytics beyond a basic level is more appropriate at the master's level.
My hypothesis is that ``Intro to DS'' is a more natural and practical introduction to computer science than most of the current intro courses, which often still rely on the same old sorting puzzles. Based on my experience with the class I created at Brown University, ``Intro to DS'' can survey many different areas of computer science and convey the core ideas behind them, while providing a major overview and understanding of how the different areas are related. I would argue that it is the ideal introductory class, as it allows students to learn about fundamentals while also being exposed to cutting-edge research and real-world use cases.
The University of Washington offers transcriptable options for many graduate programs that enhance a domain graduate degree, for example in Chemical Engineering, by having the student take classes from the core competency list (Machine Learning, Stats, Visualization and Data Management). This is a great time domain-based students to be exposed to the ideas of data science. They have achieved enough education in their domain area and are now ready to incorporate data science. However, there is nothing preventing optional exposure to undergrads to the data science competencies. In fact, to be competitive in graduate school admins, undergraduates should be avail themselves of any data science opportunities at their university For methodological domains, e.g. Computer Science, Statistics, Applied Math, for example, the undergraduate environment is a great place to have data science components as part of their education. In short, undergrads should have exposure to ideas and concepts but mastery is expected at the graduate level.
Some principles of data science can be taught at the undergraduate level but a working data scientist should have graduate education in data science. It requires a certain level of maturity and breadth of knowledge that likely will not fit into an undergraduate curriculum.
The competencies identified in the first question require a level of intellectual maturity that may not be associated with the "typical" undergraduate. However, introduction at the undergraduate level is certainly possible with depth being acquired through additional study either through formal or informal continuing education. It seems that introduction of the competencies is necessary at the undergraduate level but is not sufficient.
Certification is also an option, for some areas. See the table below for more. Certification is suitable for entry level data analyst capable of daily operational activities, basic data acquisition activities, data exploration activities and monitoring duties related to production content. A data scientist would have an undergrad degree biased in data management, and would be fully capable of delivery activities defined below but perhaps not ready to manage others throughout that process. Necessary management skills would arrive at the Masters level introducing a Managing Data Scientist who can not only execute in the defined process, but manage others in it. Growing beyond the process, setting an overall organization direction likely requires a PhD in Machine

Learning perhaps applied to a specific industry setting: the chief data scientist.



Currently, there is need for both. Today's undergraduates in these areas are very motivated and willing to take on challenging material. As such, they can, with proper guidance, handle the programming, mathematics, communication, and critical thinking challenges associated with these areas. It would be a disservice to not make available this material for a properly motivated undergraduate student, since these were be very important skills for a career in this area.

However, many students who already have a bachelor's degree (even those with ones in CS, Math, Stats, or some other heavily quantitative field) from more than 5 years probably did not have access this material. Still many universities do not offer programs with this sort of emphasis for undergraduate students. Hence, the need is still (and will continue to be for some time) appropriate for graduate programs.

Moreover, I anticipate in the future (maybe 10 years from now) that most undergraduates at top universities interested in data science will be able to get a serviceable training in data science. At this point, I believe there will still be a place for graduate education in this area, but these programs will be able to focus on more specialized areas within data science, for instance advanced machine learning, or advanced distributed systems and data management.

Both levels are appropriate. Each data science challenge will require teams of people drilled down in different areas and to different levels. The problem will drive the configuration of the team. So training people on both levels is important.

Both are appropriate. In the federated model, there are needs for general competencies that are well suited to an undergraduate education; there are more sophisticated competencies for the other archetypes.
This surely varies by the kind of institution and its ambitions. I don't see why a data science degree program that is at its core business analytics couldn't be offered at many undergraduate institutions. I do believe that a data science degree program (major, e.g., leading to a B.S.) that aims beyond that would need to be carefully designed at the undergraduate level. Masters' programs are in a sense a different beast, as they are often better thought of as a program of advanced training directed mainly to an economic/job market niche, rather than having the broader ambitions that some institutions associate with an undergraduate major.
Both are appropriate. Students can obtain an undergraduate or graduate degree in computer science, statistics, science, psychology, etc., so why should data science be any different? Clearly, a master's degree would offer a deeper, more theoretical underpinning.
I think there is a need for programs at both levels. Data science is learned and honed over time. Across concepts, three are foundations as well as sophisticated nuances, which can be taught at undergraduate and graduate levels, respectively.

Question 3. Many computing disciplines have developed a body of knowledge either as part of a curriculum development project (such as the ACM/IEEE and ACM/AIS initiatives for Computer Engineering, Computer Science, Information Systems, Information Technology and Software Engineering) or separately (such as SWEBOK). If you were considering a similar effort for data science, what would be your top three to five choices to be included as top level categories (Knowledge Areas)? If possible, please describe each area with a couple of sentences. If you want to include more than five Knowledge Areas, it is perfectly fine, but please indicate the top five.

1. Computing Competencies
2. Mathematical Foundations—to include statistics, optimization, and decision sciences
3. Modeling Foundations—understanding of how to construct mathematical models of real world situations and extract insights from them
4. Business Foundations—leadership, organizational change, communicating technical information to a non-technical audience

A note: even though 2 and 3 may be more often associated with business analytics, I think it is important for data scientists to have an understanding of how the data will be used.

I am not 100% sure, what level these knowledge areas should be as I am not very familiar with SWEBOK etc. On a high-level:

1) Data Cleaning / Data integration techniques (most of the time goes there)

2) Big Data Management tools (often being able to analyze more data is better than applying sophisticated and highly complicated techniques, as shown in many use cases)

3) Machine Learning (obviously)

4) Statistics (I strongly believe, statistics are immensely important. However, Data Science is not about statistics alone. Often very simply and potentially compute techniques, such as permutation tests, suffice).

5) Visualization (communicating the final result is extremely important)

6) Every data science curriculum should also train the student in one or two domains, such as business, biology, chemistry, … In the end, the person has to speak the domain knowledge.

At the UW, we have built our programs around the following four knowledge areas: Statistics, Machine Learning, Data Management (Databases), and Visualization. I personally believe strongly that basic Software Engineering is an additional knowledge area. Most of machine learning and visualization have a substantial statistics component. Knowledge of basic statistics, parameter estimation, and the underlying calculus is important to provided a baseline for the other competencies. Machine learning is the expression of statistics and what appears to be most important in the education is practical exposure to real world learning problems. The goal is to provide students the abilities to choose the correct machine learning technique (given pitfalls, etc.) for a given application. In terms of data management, this is really about databases with some pieces for managing large datasets, understanding privacy and ethics related to large data sets, usually health related. Visualization, like machine learning, is very much about exposing the students to various visualization environments (e.g. D3), various design patters in presenting

information, and giving students an understanding of how visualization is informed by databases that hold the data and machine learning methods used to extract knowledge that is being visualized. Finally, software engineering in some form is essential for domain based data scientists. We are finding that learning programming languages is not the impediment it once was, given the abundance of learning resources on the internet and more streamlined languages like python. What appears to be problematic for domain based data scientists is learning how to write good code that is reusable, reproducible, documented, and tested. To that end, a lightweight software engineering class is encouraged for undergraduate and graduate domain based data scientists.

Programming; Data Shaping; Statistical Inference; Predictive Modeling; Information Presentation and Visualization

The knowledge areas are (1) Data Management: This includes data acquisition, data munging (wrangling), and data manipulation. This applies regardless of the size of the data set and requires knowledge of SQL and a programming language (R or Python as current examples. (2) Statistical Inference and Machine Learning: This describes the ability to draw conclusions from data and build systems that improve with experience. (3) Software Engineering: This describes the ability not only to program but to create software products that are reliable, and the ability to adapt to computing environments, e.g. parallel or distributed computing. (4) Visualization: The ability to represent data visually to derive insight. (5) Modeling and Simulation: The ability to describe problems, both simple and complex, mathematically so that computational tools may be utilized to derive insight.

Although many possible paths exist to becoming a chief data scientist, the path rooted in data management is likely the path of least resistance. In terms of knowledge areas, four key areas from differing faculties are outlined below:

Role	**Knowledge Areas**			
	Mathematics	**Computer Science**	**Data Management**	**Business Management**
Data Analyst (Certificate)	• Statistics (Descriptive, Regression, Random Variables, CLT, Chi)	• Programming (R, Hadoop, Python, JSON, XML, DataFrames, Spark, Knime,) • Analysis Tools • Big Data Architectures (Distributed systems, 3Vs, social intelligence, IOT, Sensors, Smart Grid)	• Data Storage & Operations (Relational, NoSQL, CAP, Sharding)	• Communication • Decision Modeling
Data Scientist	• Statistics II (p-test, Bayes,	• Computer Systems (O(n), Hash,	• Data Integration • Data	• **Journalism* • Business Strategy

(Undergrad)	Distributions, ANOVA+, PDFs, Covar+Correl • Linear & Matrix Algebra	BTree) • Data Modeling • Visualization Tools & Techniques	Warehousing & Business Intelligence (OLAP, MOLAP,) • Data Security	• Performance Management • Ethics • Privacy
Managing Scientist (Masters)	• C&O (network theory, linear programming) • Predictive Analytics (timeseries, ARIMA, Trees, k-Nearest, Class+Pred,	• Computer Architecture • Network Analysis	• Data Quality	• "Industry" Analytics • Consulting Skills • International Data Policies
Chief Data Scientist (Doctoral)	• Supervised Learning (Naïve, Bayes, Caret, Mahout, Kernel Density) • Prescriptive Analytics	• Enterprise Architecture • Learning Systems	• Data Governance	• Leadership Skills • High Performance Teams • Descriptive Analytics • "Industry" Specialization • Legislating Data Privacy

1. Machine Learning and Data Mining: How can computers make sense of data with some autonomy (saying "fully automatically" would be overstating what is possible). The boundary between these names is somewhat fluid (and could possible also include the nomenclature "computational statistics"). I view machine learning as using (partially annotated data) to make informed or guided decisions and data mining as how to convert the data so it can be efficiently analyzed (or learned from) in a way that preserves meaning.

2. Computational Efficiency: This includes understanding the bottlenecks into processing data efficiently and scalably. This is typically taught in classes on Advanced Algorithms and Databases. The algorithms side is in the abstract and mathematical design side, and also includes topics such as advanced data structures, often randomized ones. The database sides focuses more on specifics of entrenched systems such as SQL, NoSQL, and the Hadoop-Spark ecosystem.

3. Experimental Design: How to design proper experiments, and run them at scale. You have a new analysis technique or new algorithm, how to evaluate if this is actually better than existing ones. This involves both understanding of basic probability and statistics, but also the scripting and systems tools so you can do this efficiently, and perhaps at scale.

4. Visualization and Communication: With complex data, generally understood (outside of academia) statistics may not be capable of simply capturing what is happening in that data. Visualization tools can go beyond this; however, one needs to have proper training so what is conveyed does not distort the truth. Speaking, writing, and presentation skills are also essential in this space.

I believe it may be best not to have a single model program right now, rather broad suggestions. At this stage, it is exciting to see how various programs develop and produce. Creating an accreditation requirement would stifle this innovation, and limit the potential of the field as its needs adapt.

Moreover, many different current programs have very different focuses and strengths. One housed in a CS department will be far different than one in a business school. This is true not only in what can be expected of the incoming students, but also of which market the students are being trained to serve. I imagine the same is true with programs with emphasis in various areas such as economics or health care or any scientific or engineering discipline.

• Data collection, archival, and access
• Data visualization
• Predictive and data analytics, including
  o Machine learning and data mining
  o Computational Statistics (statistical learning)
• High Performance Computing
• Social and ethical aspects of data

I feel the EU's 'European Data Science Academy' has done a good job in identifying the main areas: http://edsa-project.eu/resources/curriculum/ Of the 15 modules, the top five from my perspective are Statistical/Mathematical Foundations, Data Management and Curation, Distributed Computing, Machine Learning/Data Mining/Basic Analytics and Process Mining.
Leaving this to colleagues in the computing disciplines.
1) Data and algorithms, 2) big data, 3) statistical thinking (including bias, confounding, summary statistics, and the basics of inference), 4) modelling, data mining, or supervised learning, and 5) unsupervised/machine/statistical learning. In addition, 6) some cognate field to develop subject matter expertise would be beneficial, which could include more theoretical underpinning in statistics or computing.
Data collection and manipulation: This includes the acquisition and manipulation of data. E.g., data management, data warehousing, query languages (SQL/No SQL) Data analysis: This includes the use of statistics and artificial intelligence to solve problems. E.g., statistics – descriptive and discovery, artificial intelligence, software engineering Problem solving: This includes the application of data science to solve meaningful problems. E.g. critical thinking, problem solving, requirements elicitation, reporting and visualization, management communication, domain area, and process management, change management Data ethics, security, and controls: This includes practices and approaches to ensure that data science is applied appropriately and falls within legal, regulatory, and ethical boundaries.

Question 4: Please list up to three "must read" references to your own or others' works related to data science education or data science in general. For each, please include a 1-2 sentence annotation

explaining why others "must read" it. If you can send us an electronic copy of these readings, we will make them available on a password-protected webpage for workshop participants.

I think we are close to being able to develop a model *core* curriculum, although the home department and faculty expertise of individual programs will do much to influence how that core is taught, what aspects of data science and analytics are emphasized, and what electives are available to students. I hesitate to push anything resembling "accreditation", as I think that can tend to lend focus more towards satisfying criteria than towards innovation in our curricula. I think a better approach is certification. If programs can produce students that can achieve a standard certification in data science in analytics then that can serve as evidence that the program is effectively teaching students appropriate core skills, in whatever way they choose to do so.
I believe we do. As a first step we created the Data Science Teaching Initiative at Brown University: a place to share teaching material between lecturers. This platform will go online in October and, if wanted, I could give a demo at the workshop.
I believe we do and the UW Data Science transcriptable option is our version of this. Please see the web site that describes it: http://lazowska.cs.washington.edu/ADS.pdf
I believe they would; it would force educators to think about the truly necessary core skills and the different practice areas built on that. It would also lead to an expectation of what a data scientist would know when they graduate from and undergraduate or graduate program.
Models are important and help ensure an efficient use of resources. Rather than a model curriculum a statement on the required competencies with successful implementation examples might be a more productive approach. It seems unlikely that a "one size fits all" curriculum is possible especially at this in the development of the field so a focus on the required competencies would allow institutions flexibility while ensuring students acquired the necessary skills and knowledge. These resources would be useful in encouraging quality program development, encouraging research, and identifying funding opportunities.
The technology components are certainly beginning to crystalize with industry standard communication well in place (TCP/IP) and now storage is settling on Hadoop with its native ability to query content in place. The Open Source movement propelled R to a global audience and its adoption rate is very high and steadily increasing. These elements combined: communication, storage, and computational processing are now a commodity. Similarly in the business world, emphasis on corporate performance management, the instrumentation and measurement of business attributes, has produced a commoditized bed of key performance metrics. The beauty of this is the even split in the business intelligence domain that requires a certain degree of maturity to explore and sustain these metrics and the world of opportunities it offers. Performance metrics once embraced by an organization compel forecasting, trending and predictive analytics: the bread and butter of today's data scientists. And these propel the need for further learning systems, a requirement for today's chief data scientists. The corporate performance management model works in commercial settings as well as not-for-profit and public sector as well as educational institutions as well: if it's not being measured, then it will not be managed.
Yes, many of the needed skills are already mature fields. It is time to take a stab at this, but to create something that can grow with the field. That is, whatever is created should be extensible

and modifiable.
Yes. I think many universities are struggling to start programs and/or define them. Model curriculum would be helpful to those in the trenches.
One way to find out would be to try to create a set of model curricula – thinking of this effort less as a curriculum design process itself, and more as a way to get different opinions and approaches into view. In practice, at our institution (Berkeley) we have been looking at other universities' programs to see what they're experimenting with. Our reaction has not been to take over what we're seeing elsewhere, but try to build it on foundations from the freshman year forward. My sense is it's too early to begin an effort around accreditation criteria, if that's meant as an intellectual definition of a terrain that's still in huge flux. On the other hand, if an effort around accreditation criteria is intended (as it usually is, observed sociologically) to draw boundaries around a field in order to keep out pretenders and charlatans, then there could be value in that. One senses a lot of anxiety, among both experts and the lay public, about what circulates as "data-driven" insight and how well it is grounded behind the screen of technical expertise.
Yes, there are plenty of exemplar programs in universities and in organizations from which to learn and develop model curriculum and accreditation criteria. Such resources would absolutely aid in the field's maturation. I believe academics and practitioners alike would welcome such resources.

Question 5: Please list up to three "must read" references to your own or others' works related to data science education or data science in general. For each, please include a 1-2 sentence annotation explaining why others "must read" it. If you can send us an electronic copy of these readings, we will make them available on a password-protected webpage for workshop participants.

Michael F. Gorman, Ronald K. Klimberg (2014) Benchmarking Academic Programs in Business Analytics, *Interfaces*, 44(3):329-341.

> The authors surveyed a representative selection of business analytics programs in the US to understand entry requirements, topics covered, and job prospects for graduates. Their survey revealed a number of commonalities as well as several novel developments.

Fourth Paradigm

> A tour de force of data science (before it was called data science) across disciplines from oceanography to astronomy to healthcare. Then it goes into the key methodologies. Describes how we got to the fourth paradigm of science.
> *Available for free from Microsoft: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf*

A list of best hits from data science: http://arxiv.org/abs/1503.08776

> *Chris and Mark solicited lists of influential works in data driven discovery and this paper outlines the results. Abstract:*
>
> *The Gordon and Betty Moore Foundation ran an Investigator Competition as part of its Data-Driven Discovery Initiative in 2014. We received about 1,100 applications and each applicant had the opportunity to list up to five influential works in the general field of "Big Data" for scientific discovery. We collected nearly 5,000 references and 53 works were cited at least six times. This paper contains our preliminary findings.*

Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society

> This paper written by a working group of the American Statistical Association provides an excellent overview of the fields being impacted by "big data" and identifies the need for a multi-disciplinary approach to research.

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2e

> This book informs the knowledge areas described previously.

Data Science in Statistics Curricula: Preparing Students to Think with Data:

> Hardin, Hoerl, Horton, and Nolan: This paper provides sound advice on how to integrate data science into curricula.

Tidy Data (Wickham):

> This paper is important for understanding the critical data management component.

DAMA DMBOK v2

The Elements of Statistical Learning - Data Mining, Inference and Prediction

Business: Strategy, Development, Application

Rethinking Abstractions for Big Data: Why, Where, How, and What | http://arxiv.org/abs/1306.3295

Data Driven, by D J Patil and Hilary Mason

Data: Emerging Trends and Technologies

How sensors, fast networks, AI, and distributed computing are affecting the data landscape

by Alistair Croll, O'Reilly

http://www.oreilly.com/data/free/data-emerging-trends-and-technologies.csp

[free but must enter email address to download]

Probably the best overall snapshot of the state of the art.

Algorithms and Bias: Q. and A. With Cynthia Dwork

by Claire Cain Miller, *New York Times*

http://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html

Explains some of the ethical issues inherent in machine learning and the use of algorithms.

Big Data and Its Exclusions

by Jonas Lerman, *Stanford Law Review*

http://www.stanfordlawreview.org/online/privacy-and-big-data/big-data-and-its-exclusions

Examines issues of privacy from a novel perspective, arguing that being excluded as a group from big data collection can be just as damaging as being included too personally.

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

by Steve Lohr, *New York Times*

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Explains the "iceburg under the water" of data science: how the majority of the work is still in hand-cleaning and combining data sets until they are ready for the part that gets all the media attention: the algorithms and insights.

== ALSO INSIGHTFUL ==

2014 Data Science Salary Survey:

Tools, Trends, What Pays (and What Doesn't) for Data Professionals

by John King and Roger Magoulas, O'Reilly Media

http://www.oreilly.com/data/free/2014-data-science-salary-survey.csp

[free but must enter email address to download]

Looks at popular tools, salary data, and the connections between the two. Could be useful for deciding which tools to teach and how to discuss job placement prospects or a program's potential ROI for students.

Gillian Tett gets it very wrong on racial profiling

by Cathy O'Neil

http://mathbabe.org/2014/08/25/gilian-tett-gets-it-very-wrong-on-racial-profiling/

Gets at the same ethical issues as the Dwork interview, only a bit more pointedly, directly, and succinctly. O'Neil is an "algorithm auditor" and is currently at work on a book for Random House to be called *Weapons of Math Destruction*, about the damage done by biased algorithms in areas such as finance, education, and the criminal justice system.

The Hot Spotters:

Can we lower medical costs by giving the neediest patients better care?

by Atul Gawande, *The New Yorker*

http://www.newyorker.com/magazine/2011/01/24/the-hot-spotters

This piece is a few years old now, and a bit long, but still one of the best pieces I've read about the possibilities of big data in healthcare. Provides a good perspective and a memorable story.

Gupta, Babita; Goul, Michael; and Dinter, Barbara (2015) "Business Intelligence and Big Data in Higher Education: Status of a Multi- Year Model Curriculum Development E:ort for Business School Undergraduates, MS Graduates, and MBAs," *Communications of the Association for Information Systems*: Vol. 36, Article 23.Available at: http://aisel.aisnet.org/cais/vol36/iss1/23

Early work in model curriculum development – provides good background to topic.

Chapters by Michael Stonebraker and Tom Davenport, Getting Data Right, O'Reilly, preview edition compliments of tamr, 2015

Quick read that has served to reshape my thinking about the federated model and the curriculum archetypes.

"Data Sciences @ Berkeley: The Undergraduate Experience," Sketch 1.2, by the Data Sciences Education Rapid Action Team (Version: 1/19/2015),

http://ls.berkeley.edu/about-college/l-s-divisions/undergraduate-division/DataScienceCurriculumSketch.pdf

The "sketch" of the program that is beginning this year at Berkeley. It starts with a set of freshman offerings (piloting in Fall 2015 with "Foundations of Data Science" plus a range of "connectors") and will build in future years toward a minor and major. More details (links to syllabi, etc.) at http://data.berkeley.edu/data-science-education-program

J. Hardin, R. Hoerl, N. J. Horton, and D. Nolan, "Data science in the statistics curricula: Preparing students to 'think with data,'" 7/24/2015, http://arxiv.org/ftp/arxiv/papers/1410/1410.3127.pdf

Even as data science is being pushed forward in part by computing/CS programs, statistics educators have been grappling with it for a long time.

Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society

A Report from the Working Group of the American Statistical Association

www.amstat.org/policy/pdfs/Big**DataStatistics**June2014.pdf

I like some of the writings of Randy Bartlett (who focuses on data science in business, i.e. business analytics). For example…

http://www.kdnuggets.com/2015/05/applied-statistics-thinking-not-toolbox.html

Wixom, B. H., T. Ariyachandra, D. Douglas, M. Goul, B. Gupta, L. Iyer, U. Kulkarni, J. G. Mooney, G. Phillips-Wren, and O. Turetken. "The Current State of Business Intelligence in Academia: The Arrival of Big Data," *Communications of the AIS*, 34, 1 (2014). http://aisel.aisnet.org/cais/vol34/iss1/1

This is the third of three Business Analytics Congress reports (2009, 2010, 2012), all of which were created to capture current needs and trends in BI and analytics from the perspective of employers, students, and faculty. The reports are based on a survey and a face-to-face event.

Appendix B: Workshop agenda

Workshop on Data Science Education

Funded by the National Science Foundation (Award# 154513)

October 1-3, 2015

Embassy Suites Crystal City, Arlington, VA

Co-organizers: Boots Cassel (Villanova University) and Heikki Topi (Bentley University)

Schedule of Activities

October 1, 2015 (Thursday)

7pm – 8:30pm Room: Capitol Hill	**Opening reception** for informal meet and greet **Sponsored by the ACM Education Board**

October 2, 2015 (Friday)

Unless otherwise specified, the events will take place in the Adams Morgan room. Please note that breakfast is included in the room rate and will be available starting at 6am.

8:15am – 8:30am	**Welcome, introductions, setting the stage**
8:30am – 10:30am	**Session 1: What is Data Science?** Core questions • We do not have to agree and we don't expect the group to agree regarding the fundamental identity of Data Science. It is, however, important for us as a group to set the stage by understanding and articulating the richness of perspectives on what Data Science is. The purpose of this session is to explore how many different ideas of Data Science do we, as a group, have. We would also like to find out whether or not we can bring the perspectives together into a reasonably small set.

	• Where do the divisions show up? Is it based on the initial academic discipline (CS, IS, Statistics, Information Science), the application domain (business, sciences, humanities, etc.) or the environment in which Data Science is practiced (academic, research lab, business, etc.) Process • Four prepared participant contributions (Cathryn Carson, Lise Getoor, Michael Goul, and Michael Posner) of max seven minutes (timed – 30 minutes total) • Small group discussion (45 minutes) • Report-outs and general discussion (45 minutes)
10:30 am – 10:45am	**Break**
10:45am – 12:45pm	**Session 2: What are the core knowledge and skills areas of Data Science practice?** Core questions • What are the absolutely necessary knowledge and skills areas forming the core of Data Science practice? (From the proposal: What are the specific theories and concepts that provide needed tools and techniques for those faced with mountains of data and needing to know what it can tell them? Whether it is genetics, or astronomy or social science data, or survey results, or customer behavior, what approaches will yield the knowledge needed to make decisions, design products, or set new directions for study?) • What are the special areas that apply to your department/group within your own university? Do you believe there are forms of Data Science outside your own unit's definition? Process • Brief opening presentation by Boots and Heikki based on the participant pre-workshop submissions (15 minutes) • Participants will work in small groups towards a consensus regarding the questions above (60 minutes) • Report-out and general discussion (45 minutes)
12:45pm – 1:30pm	**Working lunch**
1:30pm – 3:00pm	**Session 3: Relationship of Data Science with application domains**

	Core questions • How much of Data Science is about the theory and concepts related to handling data, independent of the application domain? How much should students of data science learn about data integrity, privacy, security, storage and manipulation, sharing, distributed processing, machine learning, statistical methods, algorithms and data structure, etc. without caring what purpose the data serves? Is it meaningful to have a Data Science degree without an application domain? How do we determine the balance between various disciplinary perspectives? • Do the answers vary depending on context, environment? Are there things we can agree on? Process • Four prepared participant contributions (Paul Anderson, David Beck, Nick Horton, and Barbara Wixom) of max seven minutes (timed ~ 30 minutes) • Small group discussions (30 minutes) • Report-outs and general discussion (30 minutes)
3:00pm – 3:15pm	**Break**
3:15pm – 5:00pm	**Session 4: Implications and potential consequences of data science practice** Core questions • Can we teach aspiring data scientists to understand implications and potential consequences of their work? Should ethics of Data Science be included in Data Science programs? Process • The session starts with a brief introduction to the topic (15 minutes) • The participants will be divided into small groups. The group process will start with a brief framing exercises in which the groups will be discussing a short case study that describes an ethical dilemma related to the use of data. Building on this experience, the groups will explore the questions specified above. (60 minutes) • The session will conclude with report-outs and discussion. (30 minutes)
5:00pm – 5:30pm	**Summary and wrap-up of the day**

	Preparation for Day 2
6:30pm – Room: Capitol Hill	**Workshop Dinner** **Sponsored by ACM Education Board,** **Bentley University, and Villanova University**

October 3, 2015 (Saturday)

Unless otherwise specified, the events will take place in the Adams Morgan room. Please note that breakfast is included in the room rate and will be available starting at 7am.

8:45am – 9:00am	**Review of Day 1** Process At the end of Day 1, the participants will be given an assignment to identify and submit a description of 1-2 key personal discoveries based on the first day. These descriptions are due by 8am on Day 2. Boots and Heikki will summarize the findings and present their analysis in order to set the stage for Day 2.
9:00am – 10:15am	**Session 5: Developing a Workforce for Data Science** Core questions • What are the most essential steps that need to be taken to develop a competent workforce in Data Science? • What types of programs in Data Science do we need? • Who should offer these programs? • What roles can professional societies play? Process • Discussion within the entire group
10:15am – 10:30am	**Break**
10:30am – 12:00pm	**Session 6: What actions are needed to move Data Science**

	education forward? Core questions • Is there enough agreement about at least some needs to suggest a curriculum recommendation effort? • Is it time to consider an accreditation effort? • Who could lead efforts in curriculum design? • Who could accredit programs in Data Science? Process • Start with a vote • Share results • Open discussion • Vote again
Noon – Adjourn	**Discussion will continue over lunch for those who are able to stay**