Data Science: A Comprehensive Overview

LONGBING CAO, University of Technology Sydney, Australia

The 21st century has ushered in the age of big data and data economy, in which *data DNA*, which carries important knowledge, insights, and potential, has become an intrinsic constituent of all data-based organisms. An appropriate understanding of data DNA and its organisms relies on the new field of *data science* and its keystone, *analytics*. Although it is widely debated whether big data is only hype and buzz, and data science is still in a very early phase, significant challenges and opportunities are emerging or have been inspired by the research, innovation, business, profession, and education of data science. This article provides a comprehensive survey and tutorial of the fundamental aspects of data science: the evolution from data analysis to data science, the data science concepts, a big picture of the era of data science, the major challenges and directions in data innovation, the nature of data analytics, new industrialization and service opportunities in the data economy, the profession and competency of data education, and the future of data science. This article is the first in the field to draw a comprehensive big picture, in addition to offering rich observations, lessons, and thinking about data science and analytics.

CCS Concepts: • **General and reference → Surveys and overviews;**

Additional Key Words and Phrases: Big data, data analysis, data analytics, advanced analytics, big data analytics, data science, data engineering, data scientist, statistics, computing, informatics, data DNA, data innovation, data economy, data industry, data service, data profession, data education

ACM Reference Format:
Longbing Cao. 2017. Data science: A comprehensive overview. ACM Comput. Surv. 50, 3, Article 43 (June 2017), 42 pages.
DOI: http://dx.doi.org/10.1145/3076253

1. INTRODUCTION

We are living in the age of big data, advanced analytics, and data science. The trend of "big data growth" [Laney 2001; CSC 2012; Beyer and Laney 2012; McKinsey 2011; Vesset et al. 2012], or "data deluge" [Hey and Trefethen 2003], has not only triggered tremendous hype and buzz, but more importantly presents enormous challenges that in turn bring incredible innovation and economic opportunities. Big data has attracted intensive and growing attention, initially from giant private data-oriented enterprise and lately from major governmental organizations and academic institutions. Typical examples include large data-centric projects in Google, Facebook, and IBM, and strategic initiatives in the United Nations [UN 2010; USNSF 2012], EU [Commission 2014], and China [Government 2015].

From the disciplinary development perspective, recognition of the significant challenges, opportunities, and values of big data is fundamentally reshaping the traditional

This work is partially supported by the Australian Research Council Discovery Grant, under grant No. DP130102691.
Author's address: L. Cao, Faculty of Engineering and IT, UTS Advanced Analytics Institute, University of Technology Sydney, PO Box 123 Broadway, New South Wales 2007, Australia.

ACM 0360-0300/2017/06-ART43 $15.00
DOI: http://dx.doi.org/10.1145/3076253

data-oriented scientific and engineering fields. It is also reshaping those nontraditional data engineering domains such as social science, business, and management [Yiu 2012; Labrinidis and Jagadish 2012; Chen et al. 2012; Khan et al. 2014]. This reshaping and paradigm shifting is driven not just by data itself but all other aspects that could be created, transformed, and/or adjusted by understanding, exploring, and utilizing data.

The preceding trend and its potential have triggered new debate about data-intensive scientific discovery as a new paradigm, the so-called "fourth science paradigm," which unifies experiment, theory, and computation (corresponding to "empirical" or "experimental," "theoretical," and "computational" science) [Gray 2007; Hey et al. 2009]. Data is regarded as the new Intel Inside [O'Reilly 2005], or the new oil and strategic asset, and drives or even determines the future of science, technology, the economy, and possibly everything in our world today and tomorrow.

In 2005 in Sydney, we were asked a critical question at a brainstorming meeting about data science and data analytics by several local industry representatives from major analytics software vendors: "Information science has been there for so long, why do we need data science?" Related fundamental questions often discussed in the community include "What is data science?" [Loukides 2012], and "Is data science old wine in new bottles?" [Agarwal and Dhar 2014]. Data science and relevant topics have become the key concern in panel discussions at conferences in statistics, data mining, and machine learning, and more recently in big data, advanced analytics, and data science. Typical topics such as "grand challenges in data science," "data-driven discovery," and "data-driven science" have frequently been visited and continue to attract wide and increasing attention and debate. These questions are mainly posted from research and disciplinary development perspectives; while there are many other important questions, such as those relating to data economy and competency, these perspectives are less well considered in the conferences referred to previously.

A fundamental trigger for the preceding questions and many others not mentioned here is the exploration of new or more complex challenges and opportunities [Jagadish et al. 2014; Cao 2010a, 2016b; Khan et al. 2014] in data science and engineering. Such challenges and opportunities apply to existing fields, including statistics and mathematics, artificial intelligence, and other relevant disciplines and domains that have never been addressed, or have not been adequately addressed, in the classic methodologies, theories, systems, tools, applications, and economy of relevant areas. Such challenges and opportunities cannot be effectively accommodated by the existing body of knowledge and capability set without the development of a new discipline. On the other hand, data science is at a very early stage and is engendering enormous hype and even bewilderment; issues and possibilities that are unique to data science and big data analytics are not clear, specific, or certain. Different views, observations, and explanations—some of them controversial—have thus emerged from a wide range of perspectives.

There is no doubt, nevertheless, that the potential of data science and analytics to enable data-driven theory, economy, and professional development is increasingly being recognized. This involves not only core disciplines such as computing, informatics, and statistics, but also the broad-based fields of business, social science, and health/medical science. Although very few people today would ask the question we were asked 10 years ago, a comprehensive and in-depth understanding of *what data science is*, and *what can be achieved with data science and analytics research, education, and economy* [Cao 2017], has yet to be commonly agreed.

Motivated by the preceding concerns and observations, this article shares the findings from a comprehensive survey of the journey from statistics and data analysis to data science. It constructs an overview of data science as a field in terms of its research, innovation, economy, profession, and education.[1] This is built on (1) our observations and experience in providing real-life data innovation and practices to large government and industry organizations; (2) the education and training opportunities that have been created for professionals at various levels; and (3) reflection on our view on critical issues, future directions, and strategic opportunities in data science and analytics.

Focusing on data science (rather than big data), it is clear that only a few articles and references have discussed its history and contents, such as in Press [2013], Donoho [2015], and Galetto [2016]. A comprehensive review of data science was provided in Donoho [2015], which focuses on the evolution of data science from statistics. To the best of our knowledge, this article is the first in the field to present such a comprehensive and in-depth survey and overview. Unlike studies that focus on evolution and specific disciplinary perspectives, this article provides an introduction to the major aspects and domains of data science research, economy, profession, disciplinary development, and education. This overview complements our other contributions on specific data science issues and perspectives, that is, the realities and pitfalls in Cao [2016c], the challenges and disciplinary directions in Cao and Fayyad [Cao 2016b], the profession and education of data science in Cao [2016d], and the book on understanding data science in Cao [2017].

There are several key terms, such as data analysis, data analytics, advanced analytics, big data, data science, deep analytics, descriptive analytics, predictive analytics, and prescriptive analytics, which are highly connected and easily confusing. Table I lists and explains them, which are also the key terms widely used in this review. A list of data science terminology is available at www.datasciences.org.

The article is organized as follows. Section 2 tracks the progression from data analysis to data science, and addresses the fundamental question: *What is data science?* In Section 3, the main features, initiatives, activities, and status of the era of data science are summarized. The evolution, state-of-the-art, paradigm shift, and major tasks of deep analytics, as the keystone of data science, are discussed in Section 4. Major challenges and directions of data-driven innovation are presented in Section 5. Section 6 summarizes new data-driven industrialization and service opportunities. The data science profession, competency, role of data scientists, and course framework are summarized in Section 7. The future of data science is briefly discussed in Section 8, followed by the conclusion of this work.

2. FROM DATA ANALYSIS TO DATA SCIENCE

This section summarizes the findings of a comprehensive survey, including ours in Cao [2016c], Cao and Fayyad [Cao 2016b, 2016d] and others such as in Press [2013], Donoho [2015], and Galetto [2016]), of the journey from data analysis to data science and the evolution of the interest in data science. Subsequently, the question "What is data science?" is addressed.

2.1. The Data Science Journey

It is likely that the first appearance of "data science" as a term in literature was in the preface to Naur's book "Concise Survey of Computer Methods" [Naur 1974] in 1974. In that preface, *data science* was defined as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences." Another term, "datalogy," had previously been introduced in 1968 as "the science of data and of data processes" [Naur 1968]. These definitions are clearly more specific than those we discuss today. However, they have

[1]Interested readers may refer to a monograph: L. Cao, Understanding Data Science, soon to be published by Springer, for comprehensive discussions about data science.

Table I. Some Key Terms in Data Science

Key terms	Description
Advanced analytics	Refers to theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insights in big data, which cannot be achieved by traditional data analysis and processing theories, technologies, tools, and processes.
Big data	Refers to data that are too large and/or complex to be effectively and/or efficiently handled by traditional data-related theories, technologies, and tools.
Data analysis	Refers to the processing of data by traditional (e.g., classic statistical, mathematical, or logical) theories, technologies, and tools for obtaining useful information and for practical purposes.
Data analytics	Refers to the theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insight into data. Data analytics consists of descriptive analytics, predictive analytics, and prescriptive analytics.
Data science	Is the science of data.
Data scientist	Refers to those people whose roles very much center on data.
Descriptive analytics	Refers to the type of data analytics that typically uses statistics to describe the data used to gain information, or for other useful purposes.
Predictive analytics	Refers to the type of data analytics that makes predictions about unknown future events and discloses the reasons behind them, typically by advanced analytics.
Prescriptive analytics	Refers to the type of data analytics that optimizes indications and recommends actions for smart decision-making.
Explicit analytics	Focuses on descriptive analytics typically by reporting, descriptive analysis, alerting, and forecasting.
Implicit analytics	Focuses on deep analytics, typically by predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery.
Deep analytics	Refers to data analytics that can acquire an in-depth understanding of why and how things have happened, are happening, or will happen, which cannot be addressed by descriptive analytics.

inspired today's significant move to the comprehensive exploration of scientific content and development.

The evolutionary journey from data analysis [Huber 2011] to data science started in the statistics and mathematics community in 1962. It was stated that "data analysis is intrinsically an empirical science" [Tukey 1962].[2] Typical original work on promoting data processing included *information processing* [Morrell 1968] and *exploratory data analysis* [Tukey 1977]. It was suggested that more emphasis needed to be placed on using data to suggest suitable hypotheses to test. This contributed to the later term of "data-driven discovery" in 1989 [KDD89 1989]. In 2001, an action plan was suggested in Cleveland [2001] that would expand the technical areas of statistics toward data science.

Playing a major role in statistics, *descriptive analytics* (also called *descriptive statistics* in the statistics community) [Stewart and McMillan 1987] quantitatively summarizes or describes the characteristics and measurements of a data sample or set. Today, descriptive analytics forms the foundation for the default analytical and reporting tasks and tools in typical data analysis and business intelligence projects and systems.

Our understanding of the roles of data analysis in those early years extended beyond data exploration and processing to the aspiration to "convert data into information and knowledge" in 1977 [IASC 1977]. More than 20 years later, this desire fostered the

[2]On this basis, David Donoho argued that data science had existed for 50 years and questioned how/whether data science really differs from statistics [Donoho 2015].

origin of the currently popular community of the ACM SIGKDD conference, specifically the first workshop on Knowledge Discovery in Databases (KDD for short) with IJCAI'1989 [KDD89 1989]. In KDD and other data mining conferences, "data-driven discovery" was adopted as one of the key themes of these events. Since then, key terms such as "data mining," "knowledge discovery" [Fayyad et al. 1996], and *data analytics* [Renae 2011] have been increasingly recognized not only in computer science but also in other areas and disciplines. *Data mining and knowledge discovery* is the process of discovering hidden and interesting knowledge from data. Today, in addition to the well-recognized events KDD, ICML, NIPS, and JSM, many regional and international conferences and workshops on data analysis and learning have been created. The latest development is the creation of global and regional conferences on data science, especially the IEEE International Conference on Data Science and Advanced Analytics [DSAA 2014]. DSAA has received joint support from IEEE, ACM, and the American Statistics Association, in addition to industry sponsorship. The preceding efforts have ostensibly made data science the fastest growing and most popular computing, statistics, and interdisciplinary communities.

The development of data mining, knowledge discovery, and machine learning, together with the original data analysis and descriptive analytics from the statistical perspective, forms the general concept of "data analytics." The initial data analysis focused on processing data. *Data analytics* is the multidisciplinary science of quantitatively and qualitatively examining data for the purpose of drawing new conclusions or insights (exploratory or predictive), or for extracting and proving (confirmatory or fact-based) hypotheses about that information for decision-making and action.

Analytics has also become more business oriented [Kohavi et al. 2002]. It now extends to a variety of data and domain-specific analytical tasks, such as business analytics, risk analytics, behavior analytics [Cao et al. 2012], social analytics, and web analytics (also generally termed "X-analytics"). Domain-specific analytics fundamentally drives the innovation and application of data science. Both domain-specific and data-specific analytics and theoretical data analytics have together formed the keystone of data science.

Figure 1 summarizes the data science journey. It presents the evolution in terms of representative moments, events and major aspects of disciplinary development, government initiatives, scientific agendas, typical socioeconomic events, and education.

2.2. Online Search Interest Trends

According to Google Trends [Google 2016d], the online search interest over time in "data science" is similar to the interest in "data analytics," but is 50%–100% less than the interest in "big data." However, the historical search interest in data science and analytics is roughly double the interest shown in big data about 10 years ago. Compared to the smooth growth of interest in data science and analytics, the interest in big data has experienced a more rapid increase since 2012. When we googled "data science," 83.8M records were returned, compared to 365M on "big data," and 81.8M on "data analytics."

Although they do not reflect the full picture, the Google search results in the last 10 years, shown in Figure 2, indicate the following: (1) Data science, data analysis, and data analytics have much richer histories and stronger disciplinary foundations than big data. (2) The significant boom in big data has been fundamentally business related, while data science has been highly linked with research and innovation. (3) Data analysis has always been a top concern, although search interest has been flattened and diversified into other hot topics, including big data, data science, and data analytics. (4) Interestingly, the word "advanced analytics" has received much less attention than all other terms, reflecting the fact that knowledge of, and interest in, more general

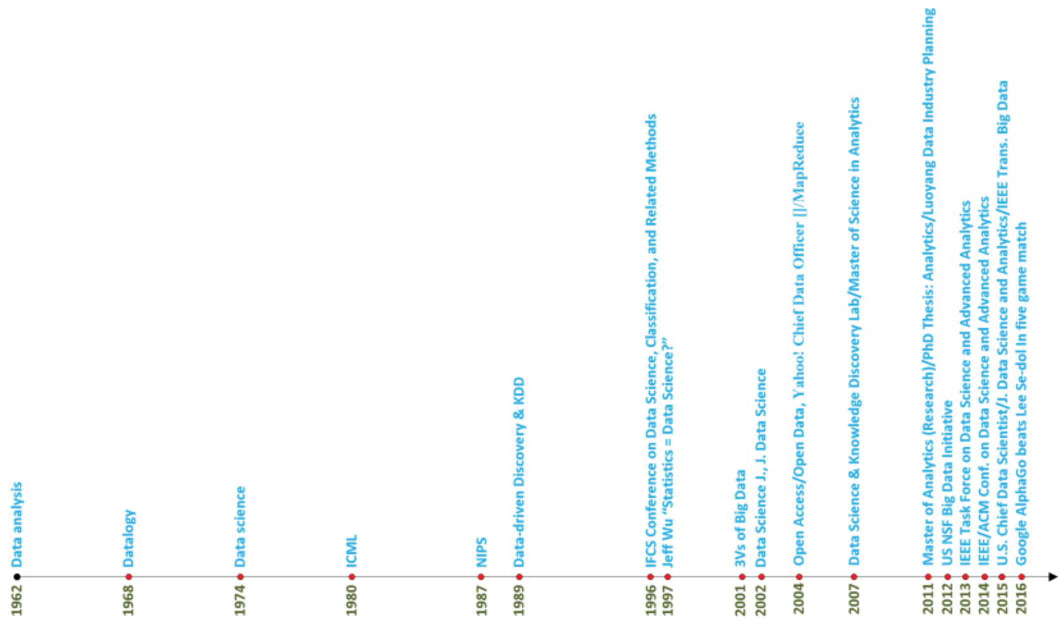

Fig. 1. Data science journey (with respect to typical events).

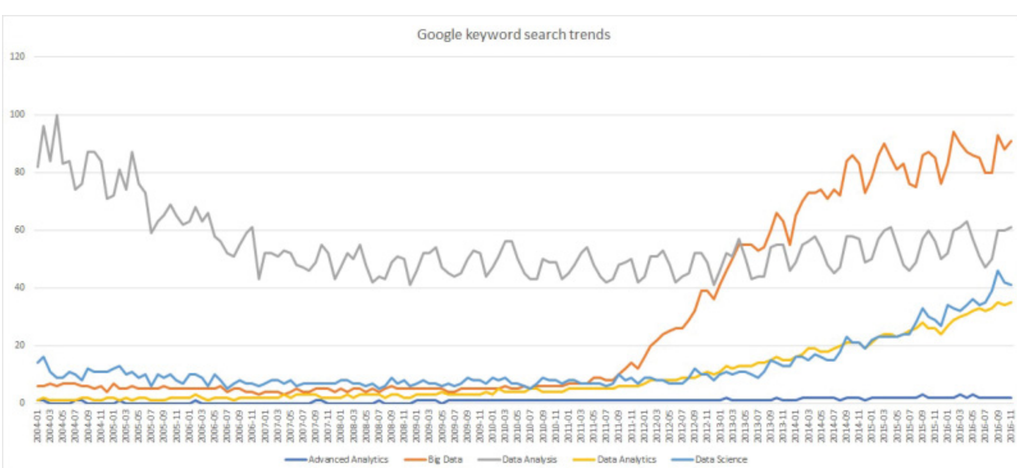

Fig. 2. Online search interest trends on data science-related keywords by Google.
Note: The data was collected on 15 November 2016.

terms like data analytics is greater than it is for more specific terms such as advanced analytics. (5) Compared to 10 years ago, scrutiny of the search trends in the past 4 years would find that big data has seen significantly increasing interest from 2012 to 2015 and then less movement; however, the interest in data science and data analytics has consistently increased, although it has grown at a much lower rate (some one-third of big data). Data analysis has maintained a relatively stable attraction to searchers during these 10 years.

2.3. What Is Data Science

The art of data science [Graham 2012] has attracted increasing interest from a wide range of domains and disciplines. Accordingly, communities or proposers from diverse backgrounds, with contrasting aspirations, have presented very different views or foci. Some examples are that data science is *the new generation of statistics*, is *a consolidation of several interdisciplinary fields*, or is *a new body of knowledge*. Data science also has implications for providing capabilities and practices for the data profession, or for generating business strategies.

Statisticians have had much to say about data science, since it is they who actually created the term "data science" and promoted the upgrading of statistics to data science as a broader discipline. This is reflected in a series of earlier actions, such as the following.

—Jeff Wu questioned in 1997 whether "Statistics = Data Science?" and suggested that statistics should be renamed "data science" and statisticians should be known as "data scientists" [Wu 1997]. The intention was to shift the focus of statistics from "data collection, modeling, analysis, problem understanding/resolving, decision making" to future directions on "large/complex data, empirical-physical approach, representation and exploitation of knowledge."
—William S. Cleveland suggested in 2001 that it would be appropriate to alter the statistics field to data science and "to enlarge the major areas of technical work of the field of statistics" by looking to computing and partnering with computer scientists [Cleveland 2001].
—Leo Breiman suggested in 2001 that it was necessary to "move away from exclusive dependence on data models (in statistics) and adopt a more diverse set of tools" such as algorithmic modeling, which treats the data mechanism as unknown [Breiman 2001].
—In 2015, a statement about the role of statistics in data science was released by a number of ASA leaders [van Dyk et al. 2015], saying that "statistics and machine learning play a central role in data science." Many other relevant discussions are available in AMSTATNEWS [ASA 2015] and IMS [Yu 2014].

A large proportion of the conceptual arguments are derived from the data-centric view. For example, data-driven science is mainly interpreted in terms of the reuse of open data [Murray-Rust 2007; OECD 2007]; data science comprises the numbers of our lives [Miller 2013]; or data science enables the creation of data products [Loukides 2011, 2012]. In Jagadish [2015], six myths were discussed: (1) size is all that matters, (2) the central challenge of big data is that of devising new computing architectures and algorithms, (3) analytics is the central problem of big data, (4) data reuse is low hanging fruit, (5) data science is the same as big data, and (6) big data is all hype. This illustrates the constituents of the ecosystem, but also shows the divided views within the communities.

Intensive discussions have taken place within the research and academic community about creating data science as an academic discipline [Smith 2006]. This involves not only statistics, but also a multidisciplinary body of knowledge that includes computing, communication, management, and decision. The concept of data science is correspondingly defined from the perspective of disciplinary and course development: for example, treating data science as a mixture of statistics, mathematics, computer science, graphic design, data mining, human-computer interaction, and information visualization [Yau 2009].

Next, we present several definitions of data science from high-level and disciplinary perspectives, building on the observations and insights we have gained from this review and relevant experience.[3]

Definition 2.1 (*Data Science*[1]). A *high-level statement* is: "data science is the science of data" or "data science is the study of data."

Definition 2.2 (*Data Science*[2]). From the *disciplinary* perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology.

Accordingly, a *discipline-based data science formula* is given as follows:

$$\begin{aligned} \textit{data science} \;=\; & \textit{statistics} + \textit{informatics} + \textit{computing} + \textit{communication} \\ & + \textit{sociology} + \textit{management} \,|\, \textit{data} + \textit{environment} + \textit{thinking}, \end{aligned} \tag{1}$$

where "|" means "conditional on."

The outputs of data science are *data products* [Loukides 2011, 2012]. We define data products next.

Definition 2.3 (*Data Products*). A data product is a deliverable from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system. The ultimate data products of value are knowledge, intelligence, wisdom, and decision.

The preceding definition of data product goes beyond technical product-based types and forms in the business and economic domain, such as social network platforms like Facebook, recommender systems like Netflix, and mobile apps like Uber. Data science enables us to explore new data-driven or data-enabled personalized, organizational, educational, ethical, societal, cultural, economic, political, cyber-physical forms, modes, paradigms, innovations, directions, and ecosystems, or even thinking, strategies, and policies. For example, there is a good possibility that large-scale data will enable and enhance the transfer of subjective autonomy to objective autonomy, beneficence, and justice in the social sciences [Fairfielda and Shteina 2014]. It can enable the discovery of indicators like Google Flu [Lazer et al. 2014] that may not be readily predicted by domain-driven hypothesis and professionals.

These platforms deliver data products in various forms, ways, channels, and domains that are fundamentally transforming our academic, industrial, governmental, and socioeconomic life and world. With the development of data science and engineering theories and technologies, new data products will be created. This creation is likely to take place at a speed and to an extent that greatly exceeds our imagination and thinking, as shown in the evolution of Internet-based products and artificial intelligence systems.

3. THE ERA OF DATA SCIENCE

In this section, we summarize the main characteristics of data science-related government initiatives, disciplinary development, economy, and profession, as well as activities in these fields, and the progress made to date. Datafication [Ayankoya et al. 2014], the Quantified Self (QS) [Swan 2013; Clay 2013; Wolf 2012; Duncan 2009; Smarr 2012;

[3]Interested readers may refer to Cao [2016c] for another definition from the process perspective.

Fawcett 2016], initiatives by governments and research institutions, and open data are discussed as the key drivers of the era of big data and data science.

3.1. Datafication and Data Quantification

Data is ubiquitous because datafication [Ayankoya et al. 2014] and data quantification are ubiquitous. In addition to the commonly seen data transactions acquired from business and operational information systems, increasingly popular and widespread datafication and data quantification systems and services are significantly strengthening the data deluge and big data realm. Such systems and services include but are not limited to wearables, Internet of Things (IoT), mobile, and social applications.

As we have seen and can predict, datafication and data quantification take place at any time and any place by anybody in any form in any way in a nontraditional manner, extent, depth, variety, and speed.

—Quantification timing: *anytime quantification*, from working to studying, day-to-day living, relaxing, enjoying entertainment, and socializing;
—Quantification places: *anyplace quantification*, from biological systems to physical, behavioral, emotional, cognitive, cyber, environmental, cultural, economic, sociological, and political systems and environments;
—Quantification bodies: *anybody quantification*, from selves to others, connected selves, exoselves [Kelly 2012] and the world, and from individuals to groups, organizations, and societies;
—Quantification forms: *anyform quantification*, from observation to drivers, from objective to subjective, from physical to philosophical, from explicit to implicit, and from qualitative to quantitative forms and aspects;
—Quantification ways: *anysource quantification*, such as sources and tools that include information systems, digitalization, sensors, surveillance and tracking systems, the IoT, mobile devices and applications, social services and network platforms, and wearable [Viseu and Suchman 2010] and Quantified Self (QS) devices and services; and
—Quantification speed: *anyspeed quantification*, from static to dynamic, from finite to infinite, and from incremental to exponential generation of data objects, sets, warehouses, lakes, and clouds.

Examples of fast developing quantification areas are the health and medical domains. We are datafying both traditional medical and health care data and "omics" data (genomics, proteomics, microbiomics, metabolomics, etc.) and increasingly overwhelming QS-based tracking data [Swan 2013] on personal, family, group, community, and/or cohort levels.

3.2. Data Initiatives by Governments

To effectively understand and utilize everywhere data, data DNA, and its potential, increasing numbers of regional and global government initiatives [Security 2015] are being introduced at different levels and on different scales in this age of big data and data science to promote data science research, innovation, funding support, policy making, industrialization, and economy. Table II summarizes the major initiatives of several countries and regions.

—The Australian Public Service Big Data Strategy [UN 2010] aims to "provide an opportunity to consider the range of opportunities presented to agencies in relation to the use of big data, and the emerging tools that allow us to better appreciate what it tells us, in the context of the potential concerns that this might raise." It addresses the identified big data strategy issues [AGIMO 2013]. Australia's whole-of-government

Table II. Government Initiatives in Big Data and Data Science

Government	Representative Initiatives
Australia	Public Service Big Data Strategy [UN 2010], Whole-of-Government Centre of Excellence on Data Analytics [AU 2016]
Canada	Capitalizing on Big Data [CA 2016]
China	Big Data Guideline [Government 2015], China Computer Federation Task Force on Big Data [CCF-BDTF 2013], China National Science Foundation big data program [CNSF 2015]
EU	Data-driven Economy [EU 2014], European Commission Horizon 2020 Big Data Private Public Partnership [Horizon 2014]
UK	UK's Big Data and Energy Efficient Computing [UK 2016]
UN	UN Global Pulse Project [UN 2010]
US	US Big Data Research Initiative [USNSF 2012], Interagency Working Group on Digital Data [CSNSTC 2009], DARPA's XDATA Program [DARPA 2016], USA NSF Big Data Research Fund [USNSF 2012]

Centre of Excellence in Data Analytics [AU 2016] coordinates relevant government activities. The Australian Research Council has granted approval to the Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers [ACEMS 2014] to conduct research on big data-based mathematical and statistical foundations. Another recent effort made by the Australian government was the establishment of Data61 [2016], which consolidated the relevant data-related human resources in the original National ICT Australia (NICTA) [NICTA 2016] and CSIRO and aims for a unified platform for data research and innovation, engagement with industry and government and academia, and software development.

—Canada's policy framework Capitalizing on Big Data [CA 2016] aims at "establishing a culture of stewardship ... coordination of stakeholder engagement ... developing capacity and future funding parameters."

—China's Guidelines [Government 2015] are aimed at boosting the development of big data research and applications, to "set up an overall coordination mechanism for big data development and application, speed up the establishment of relevant rules, and encourage cooperation between the government, enterprises and institutions." China has also set up a national strategic plan for the IoT and big data [Government 2015]. Many states and cities in China have launched national big data strategies and action plans for big data and cloud computing [CBDIO 2016; Agency 2016], such as in Beijing [Government 2016]. Probably, a very early example in China was the Luoyang City-sponsored consulting project in 2011, for which we developed a strategic plan for the City's industrial transformation to a "data industry" [Cao 2011].

—The European Union's communication Towards a Thriving Data-driven Economy [EU 2014] is "an action plan to bring about the data-driven economy of the future." It outlines "a new strategy on Big Data, supporting and accelerating the transition towards a data-driven economy in Europe. The data-driven economy will stimulate research and innovation on data while leading to more business opportunities and an increased availability of knowledge and capital, in particular for SMEs, across Europe." In 2015, the European Data Science Academy, EDSA [EU-DSA 2016] was formed.

—The United Kingdom's Big Data and Energy Efficient Computing initiative funded by the Research Councils [UK 2016] aims to "create a foundation where researchers, users and industry can work together to create enhanced opportunities for scientific discovery and development."

—The United Nations (UN) Global Pulse Project is "a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate the discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action." [UN 2010]

—The United States (US) Big Data Research Initiative [USNSF 2012] is directed toward "supporting the fundamental science and underlying infrastructure enabling the big data revolution." In 2005, the US National Science Board set the goal that it "should act to develop and mature the career path for data scientists" in its report "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" [NSB 2005]. In 2009, the Committee on Science of the National Science and Technology Council formed an Interagency Working Group on Digital Data. It published a report [CSNSTC 2009] outlining the strategy to "create a comprehensive framework of transparent, evolvable, extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data," which "will serve as a driving force for American leadership in science and in a competitive, global information society." In addition, the Defence Advanced Research Projects Agency (DARPA) launched its XDATA Program [DARPA 2016], which aims to develop computational techniques and software tools for processing and analyzing large, imperfect, and incomplete data. In 2012, the National Institute of Standards and Technology (NIST) introduced a new data science initiative [Dorr et al. 2015], and in 2013, the US National Consortium for Data Science was established [USD2D 2016].

3.3. The Scientific Agenda of Data Science

An increasing number of new scientific initiatives, activities, and programs have been created by governments, research institutions, and educational institutions to promote data science as a new field of science.

The original scientific agenda of data science has been driven by both government initiatives and academic recommendations. This was built on the strong promotion of converting statistics to data science, and blending statistics with computing science in the statistics community [Wu 1997; Cleveland 2001; Iwata 2008; Hardin et al. 2015; Hand 2015; Diggle 2015; Graham 2012; Finzer 2013]. Today, many regional and global initiatives have been taken in data science research, disciplinary development, and education, as strategic matters and agenda in the digital era. Several examples are given as follows.

—In Australia, a Go8 report [Brown 2009] suggested the incorporation of data as a keystone in K-12 education through statistics and science by such methods as creating data games for children.

—In China, the Ministry of Science and Technology very recently announced the establishment of national key labs in big data research as part of a strategic national agenda [CMIST 2016].

—In the EU, the HLSG report "Riding the Wave" [HLSG 2010] and "The Data Harvest" [HLSG 2014] urged the European Commission to implement the vision of creating "scientific e-infrastructure that supports seamless access, use, re=use, and trust of data" and foster the development of data science university programs and discipline.

—In the United States, a National Science Board report [NSB 2005] recommended that the National Science Foundation (NSF) "should evaluate in an integrated way the impact of the full portfolio of programs of outreach to students and citizens of all ages that are 'or could be' implemented through digital data collections." Different roles and responsibilities were discussed for individuals and institutions, including data authors, users, managers, and scientists as well as funding agencies. The report [CSNSTC 2009] from the US Committee on Science of the National Science and

Technology Council suggested the development of necessary knowledge and skill sets by initiating new educational programs and curricula, such as "some new specializations in data tools, infrastructures, sciences, and management."

An increasing number of research streams, strengths, and focused projects have been announced in major countries and regions, including

—The US NSF Big Data Research Fund [USNSF 2012],
—The European Commission Horizon 2020 Big Data Private Public Partnership [Horizon 2014; EU 2014], and
—The China NSF big data special fund [CNSF 2015].

Each of these supports theoretical, basic, and applied data science research and development in big data and analytics through respective scientific foundations, high-tech programs, and domain-specific funds such as heath and medical funds. Significant investment has been made to create even faster high-performance computers.

Many universities and institutions have either established or are creating research centers or institutes in data science, analytics, big data, cloud computing, and IoT. For example, in Australia, the author created the first data science lab: the Data Science and Knowledge Discovery Lab at UTS in 2007 [DSKD 2007], and the first Australian institute: the Advanced Analytics Institute [UTSAAI 2011; AGIMO 2013] in 2011 that implements the RED model of Research, Education and Development (RED) of big data analytics for many major government and business organizations. In the United States, top universities have worked on building data science initiatives, such as the Institute for Advanced Analytics at North Carolina State University in 2007 [NCSU 2007a], the Stanford Data Science Initiatives in 2014 [Stanford 2014], and the Data Science Initiatives at University of Michigan in 2015 [UMichi 2015].

3.4. Data Science Disciplinary Development

In contrast to big data that has been driven by data-oriented business and private enterprise, researchers and scientists also play a driving role in the data science agenda. Migrating from the original push in the statistics communities, various disciplines have been involved in promoting the disciplinary development of data science. This involves the disciplinary structure, intrinsic challenges and directions, course structure and curriculum design, and qualifications for next-generation data scientists [Cao 2016b, 2016c, 2016d].

In Borne et al. [2010], the authors highlight the need to train the next generation of specialists and nonspecialists to derive intelligent understanding from the increased vastness of data from the Universe, "with data science as an essential competency" in astronomy education "within two contexts: formal education and lifelong learners." The aim is to manage "a growing gap between our awareness of that information and our understanding of it." In several researches [Fox and Hendler 2014; Bailer et al. 2012; Rudin et al. 2014; Anderson et al. 2014; Baumer 2015; Bussaban and Waraporn 2015], discussions focus on the needs, variations, and addenda of data science-oriented subjects for undergraduate and postgraduate students majoring in mathematics and computing. Case studies of relevant subjects at seven institutions were introduced in Hardin et al. [2015], with the syllabi collected in Hardin [2016].

In addition to the promotion activities in core analytics disciplines such as statistics, mathematics, computing, and artificial intelligence, the extended recognition and undertaking of domain-specific data science seems to repeat the evolutionary history of the computer and computer-based applications. Data science is warmly embraced by more and more disciplines and domains in which it was traditionally irrelevant, such as law, history, and even nursing [Clancy et al. 2014]. Its core driving forces come from data-intensive and data-rich areas such as astronomy [Borne et al. 2010], neurobiology [Dierick and Gabbiani 2015], climate change [Faghmous and Kumar 2014], research assessment [Siart et al. 2015], media and entertainment [Gold et al. 2013], Supply Chain Management (SCM) [Hazena et al. 2014] and SCM predictive analytics [Schoenherr and Speier-Pero 2015], advanced hierarchical/multiscale materials [Kalidindi 2015; Gupta et al. 2015], and cyberinfrastructure [NSF 2007]. The era of data science presents significant interdisciplinary opportunities [Rudin et al. 2014], as evidenced by the transformation from traditional statistics and computing-independent research to cross-disciplinary data-driven discovery combining statistics, mathematics, computing, informatics, sociology, and management. Data science drives the disciplinary shift of Artificial Intelligence (AI) from its origins in logics, reasoning, and planning-driven machine intelligence to metasynthesizing ubiquitous X-intelligence-enabled complex intelligent systems and services [Qian 1991; Qian et al. 1993; Cao et al. 2009; Cao 2015b].

A very typical inter-, multi-, and cross-disciplinary evolutionary trend is the adoption and adaptation of data-driven discovery and science in classic disciplines from an informatics perspective. This has resulted in the phenomenon of *X-informatics* for transforming and reforming the body of knowledge. Typical examples include astroinformatics, behavior informatics [Cao 2010b; Cao and Yu 2012], bioinformatics, biostatistics, brain informatics, health informatics, medical informatics, and social informatics, to name a few [Wikipedia 2016b]. Hence, it is not surprising to see courses and subjects being offered in specific areas such as biomedical informatics, healthcare informatics, and even urban informatics.

Following the creation of the world's first coursework Master of Science in Analytics [NCSU 2007b] created at North Carolina State University in 2007, and the world's first Master of Analytics by Research and PhD in Analytics launched at the University of Technology Sydney in 2011 [UTS 2011; WIRED 2014], more than 150 universities and institutions have now either created or are planning courses in data science, big data, and analytics [Silk 2016]. The majority of these course initiatives focus on training postgraduate specialists and certificate-based trainees in business disciplines, followed by the disciplines of computer science and statistics.

Several repositories [Silk 2016; DSC 2016a; Github 2016a; Classcentral 2016; USDSC 2016] collect information about courses and subjects related to analytics, data science, information systems and management, statistics, and decision science. For example, according to DSC [2016a] and Github [2016a], there are currently about 500 general or specific subjects or courses that relate to data analytics, information processing, data mining, and machine learning, of which 78% are offered in the United States. Seventy-two percent are offered at Master's level, with only 7% at bachelor level, and 3.6% at doctoral level. About 30% are online courses. From the disciplinary perspective, some 43% of courses specifically encompass "Analytics," compared to 18% on "Data Science" and only 9% on "Statistics," Approximately 40% focus on business and social science aspects. In Classcentral [2016], 138 courses and subjects are available. A number of US programs are listed in USDSC [2016], most being created in business and management disciplines.

More than 85% of courses [DSC 2016a] cover a broad scope of big data, analytics, and data science and engineering. Some courses only offer training in very specific technical skills, capabilities, and technologies, such as artificial intelligence, data mining, predictive analytics, machine learning, visualization, business intelligence, computational modeling, cloud computing, information quality, and analytics practices. It is very rare to find courses that are dedicated to analytics project management and communication skill training [Faris et al. 2011], although several courses on decision science are offered.

Online data science courses significantly complement traditional education and typically offer a successful Internet-based data business model. The corporate training market has seen increasing competition as vendors and universities invest more resources in this area: the SAS training courses are one such example. Online courses such as those offered as a Massive Open Online Course (MOOC) and by open universities are quickly feeding the market.

Today, an increasing number of courses are offered in the MOOC mode [Fox et al. 2015; Boyer et al. 2015], such as through Class Central [Classcentral 2016; Coursera 2016], edX [Edx 2016; Udacity 2016] and Udemy [Udemy 2016]. The MOOC model is fundamentally changing the way courses are offered by utilizing online, distributed, and open data, curriculum development resources and expertise, and delivery channels and services. Course development technologies such as Google Course Builder [Google 2016c] and Open edX [OPENedX 2016] are used to create online courses and their operations.

Most of the available courses focus on classic subjects; in particular, statistics, data mining, machine learning, prediction, business intelligence, information management, and database management. New programming languages including R and Python, and cloud infrastructure MapReduce and Hadoop are highlights in these courses. Techniques related to off-the-shelf software and tools are often emphasized. Very few subjects are specified for advanced analytics, real-life analytics practices, communication, project management, and decision support. An increasing number of courses are created to address domain-specific demands, such as incorporating statistics, business analytics, web, and social network analytics into SCM predictive analytics [Schoenherr and Speier-Pero 2015].

3.5. New Data Economy and Industry Transformation

The recognition of the values and potential of data science and analytics and its rapid growth have also been driven and promoted by the evolution of a new data economy and industry transformation, such as large private data enterprise. The advancement of data science and big data analytics is conversely significantly influencing and driving the development of a new data economy, industry transformation, and increase in productivity. This wave of data economy upgrading and industry transformation features the revolution of advanced artificial intelligence-enabled technologies and businesses, and the complementary advances in AI and the AI-driven data economy are largely propelled by data science and analytics. They include inventing, commercializing, and applying infrastructures, tools, systems, services, applications, and consultations for managing, discovering, and utilizing deep data intelligence and synthesizing X-intelligences and X-complexities [Cao 2016b].

A typical indicator is the 2010 IBM Global CEO study, from which the resultant report [IBM 2010] draws the following conclusions: “Yet the emergence of advanced technologies like business analytics can help uncover previously hidden correlations and patterns, and provide greater clarity and certainty when making many business decisions.” To manage the increasing complexity, the CEOs in this study believe that “a better handle on information and mastery of analytics to predict consequences of decisions could go a long way in reducing uncertainty and in forging answers that are both swift and right.” This leads to their desire to “translate data into insight into action that creates business results” and to “take advantage of the benefits of analytics.”

Today, it can safely be said that data science has enabled the so-called “new economy” as evidenced by large private enterprises such as Facebook, Google, and Alibaba. This new *data economy* is data product-based and data technology-driven. An increasing number of organizations recognize the value of data as a strategic asset and invest in building infrastructure, resources, talent, and teams to support enterprise innovation, and to create differentiators that will lift competition and productivity. Leading Internet-based data-driven businesses [Dhar 2013], such as Google, Facebook, SAS, Alibaba, Baidu, and Tencent, have overtaken traditional enterprise giants.

Classic manufacturing-focused and core business-oriented companies, including IBM, Intel [Stonebraker et al. 2013], and Huawei, have also all launched corresponding initiatives and strategic actions for big data, IoT, and/or cloud computing, and are pursuing the strategy of data product-based transformation, productivity growth, and innovation. Data science has been their new innovation engine for productivity and competition upgrade. Core businesses, including banks, capital market firms, telecommunication service providers, and insurance companies, are leading the way in datafying, quantifying, analyzing, and using data. It is encouraging to see that other traditional business sectors, such as agriculture, tourism, retail, property, and education, are also investing in data analytics to transform their productivity and competitive advantage.

Many new start-ups and spin-offs have emerged rapidly in recent years and have focused on data-based business, products, and services. This is reflected in a fast-evolving big data landscape [BDL 2016a], which covers data sources and API, infrastructure, analytics, cross-infrastructure/analytics, open source initiatives, and applications. Every year, this changing landscape sees significant, swift growth. As a result of datafication and data quantification, new platforms, products, applications, services, and economic models such as Spark and Cloudera have quickly emerged in analytics and big data.

3.6. Data Professional Community Formation

The growth and recognition of an emerging field can be effectively measured in terms of the formation width, depth, and speed of its professional communities. The data science and analytics community is growing incredibly quickly.

The first indicator is the emergence of dedicated publication venues in this area. Several journals on data science have been established. These include the Journal of Data Science [JDS 2002], launched in 2002, which is devoted to applications of statistical methods at large; the electronic Data Science Journal [DSJ 2014] relaunched by CODATA in 2014; the EPJ Data Science [EPJDS 2012] launched in 2012; the International Journal of Data Science and Analytics (JDSA) [JDSA 2015] in 2015 by Springer; IEEE Transactions on Big Data [TOBD 2015] in 2015; and the Springer Series on Data Science [SSDS 2015] and the Data Analytics Book Series [DABS 2016].

Other publications are in development by various regional and domain-specific publishers and groups. Some examples are the International Journal of Data Science [IJDS 2016], Data Science and Engineering [DSE 2015] published on behalf of the China Computer Federation (CCF) [DSE 2015], the International Journal of Research on Data Science [IJRDS 2017], and the Journal of Finance and Data Science [JFDS 2016].

The second indicator can be found in the creation of a data science community that is significantly enhanced by conferences, workshops, and forums dedicated to the promotion of data science and analytics. There are also many well-established venues that either focus on specific aspects such as KDD and ICML or have adjusted their previous nondata and/or analytics focus, such as the traditional AI conferences IJCAI and AAAI.

—The first conference to adopt "data science" as a topic was the 1996 IFCS Conference on Data Science, Classification, and Related Methods [IFSC-96 1996], which included papers on general data analysis issues.
—The IEEE International Conference on Data Science and Advanced Analytics (DSAA) [DSAA 2014] launched in 2014, was probably the first conference series dedicated

to both data science and analytics research and practice. Cosponsored by ACM SIGKDD, IEEE CIS, and the American Statistics Association (ASA), it attracted wide and significant interest from statistics, industry, business, IT, and professional bodies. The IEEE Conference on Big Data is an event dedicated to broad areas of big data.
—Several other domain-specific and regional initiatives have emerged, such as the three initiatives in India, that is, the Indian Conference on Data Sciences, the International Conference on Big Data Analytics, and the International Conference on Data Science and Engineering.
—Several other conference series have been renamed and repositioned from their original focus on topics such as software engineering and service-based computing to connect with big data and data science, drawing mainly on key topics of interest and participants from their original areas.
—Data analytics, machine learning, and big data have eclipsed the original topics of interest in many traditionally nondata and/or analytics conferences, such as IJCAI, AAAI, VLDB, SIGMOD, and ICDE. Not surprisingly, some of these venues now frequently incorporate more than 50% of papers on data science matters.

The third indicator is the growth and development of professional (online) communities and organizations established publicly or privately to promote big data, analytics and data science research, practices and education, and interdisciplinary communications. For example:

—The IEEE Big Data Initiative [IEEEBD 2014] aims to "provide a framework for collaboration throughout IEEE," and states that "Plans are under way to capture all the different perspectives via in depth discussions, and to drive to a set of results which will define the scope and the direction for the initiative."
—The IEEE Task Force on Data Science and Advanced Analytics (TF-DSAA) [TFDSAA 2013] was launched in 2013 to promote relevant activities and community building, including the annual IEEE Conference on Data Science and Advanced Analytics.
—The International Institute of Data & Analytics [IDA 2014] aims to bridge the gaps between academia and industry through the promotion of data and analytics research, education, and development.
—The China Computer Federation Task Force on Big Data [CCF-BDTF 2013] consists of a network of representatives from academia, industry, and government, and organizes its annual big data conference with participants from industry and government.
—Several groups and initiatives promote dedicated activities of analytics and data science. For instance, Datasciences.org [2005] collects relevant information about data science research, courses, funding opportunities, professional activities, and platforms for collaborations and partnership. The Data Science Community [DSC 2016b] claims to be the European Knowledge Hub for Bigdata and Datascience. Data Science Central [DSCentral 2016] aims to be the industry's online resource for big data practitioners. The Data Science Association [DSA 2016] aims to be a "professional group offering education, professional certification, conferences and meetups" [Galetto 2016], and even offers a "Data Science Code of Professional Conduct."
—Many existing consulting and servicing organizations have adjusted their scope to cover analytics, where they previously focused on other disciplinary matters. Interdisciplinary efforts have been made to promote cross-domain and cross-disciplinary activities and growth opportunities. Examples include [INFORMS 2016], Gartner, McKinsey, Deloitte, PricewaterhouseCoopers, KPMG, and Bloomberg.

Lastly, multinational vendors, online and new economy giants, and service providers play a critical driving role in community outreach. Each of these has launched relevant initiatives, such as those by SAS [2016], IBM [2016a], Google [2016a], and Facebook [2016]. Many professional interest groups have been set up in social media, including Google groups, LinkedIn, Facebook, and Twitter, and are among the most attractive and popular venues for big data, data science, and analytics professionals to share and network.

3.7. The Open Model and Open Data

A key feature differentiating the data science era from the previous era lies in the overwhelming adoption and acceptance of the open model rather than a closed one. The *open model* enables free, distributed, and collaborative modes in every aspect of economy, society, research, and living. It supports the innovation of social media like Facebook and LinkedIn, the migration of mobile to smart phone-embedded applications, and industrial transformation such as the migration of physical shop center-based commerce to online businesses like Taobao.

Typically, open data and data sharing programs have been announced in many countries and domains, such as the US Government open data site [US-OD 2016], the UK open data project [UK-OD 2016; UK-HM 2012], the Australian Government open government data site [AU 2010, 2013] and Data Matching program [AU 1990], and the European Union Open Data Portal [EU-OD 2016] and data sharing projects [HLSG 2014]. In addition, many Open Access schemes are increasingly being accepted by academic journals.

Efforts have also been made in diverse societies to create sharable data repositories, especially for science and research. Examples of open repositories are the global climate data [Tutiempo 2016], the global terrorism database [GTD 2016], the Yahoo Finance data [Yahoo 2016], the Gene Expression Omnibus [GEO 2016], mobile data [Google 2016e], the UCI repositories for machine learning [UCI 2016], the Linguistic Data Consortium data for Natural Language Processing [LDC 2016], the TREC data for text retrieval [NIST 2015], Kaggle competition data [Kaggle 2016], and the VAST challenge [Vast 2016] for visual analytics, to name a few.

4. DATA ANALYTICS: A KEYSTONE OF DATA SCIENCE

In the age of analytics, what is to be analyzed, what constitutes the analytics spectrum for understanding data, and what form the paradigm shift of analytics takes are critical questions to be answered. We address these issues in this section.

Data and analytics form a comprehensive map that covers

- —the whole lifecycle of the data from the past to the present and the future,
- —the analytics from explicit (known) analytics and reactive understanding to implicit (unknown) analytics and proactive early prediction and intervention, and
- —the journey from data exploration (by descriptive and predictive analytics) to the delivery of actionable insights and decisions through prescriptive analytics and actionable knowledge delivery [Cao et al. 2010].

4.1. Data-to-Insight-to-Decision Whole-of-Life Analytics

As shown in Figure 3, the data-to-insight-to-decision transfer at different time periods and analytic stages is embodied along the whole-of-life analytics. This can be further represented in terms of a variety of analytics goals (G) and approaches (A) to achieve the data-to-decision goal.

- —*Past data*: the main focus of historical analytics is to explore "what happened" in the data and business, and to gain insights into "how and why it happened" through

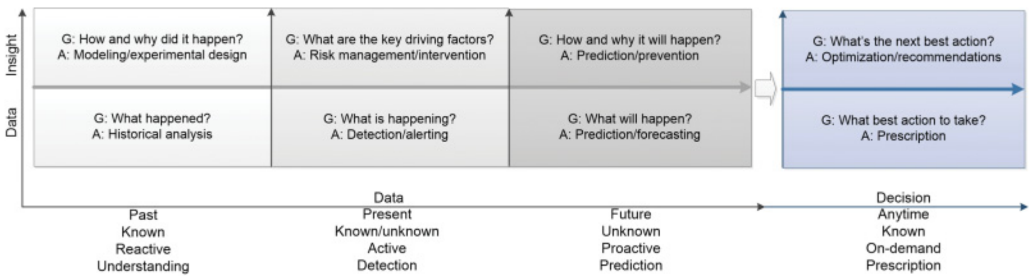

Fig. 3. Data-to-insight-to-decision whole-of-life analytics.

modeling and experimental design, etc. This stage focuses on "we know what we know" to conduct a reactive understanding of what took place.

—*Present data*: detection at this stage is mainly focused on exploring "what is happening," to generate insights about "how and why it happens." This stage addresses "we know what we do not know" with alerts generated about suspicious events, or interesting groups or patterns presented in the data and business. The insights are extracted for decision-making purposes, such as real-time risk management and intervention, to address the question "what are the key driving factors?"

—*Future data*: predictive analytics is undertaken to investigate "what will happen" in the future, and to achieve insights into "how and why it will happen" by estimating the occurrence of future events, grouping, and patterns. The aim of this stage is to solve the problem that "we do not know what we do not know" by achieving proactive understanding, forecasting and prediction, and early prevention.

—*Actionable decision*: prescriptive analytics and actionable knowledge delivery are undertaken to investigate "what best action to take" to interpret findings from the past, present, or future data. This achieves insights into "what is the next best action" and enables the corresponding optimal actions and recommendations to be undertaken based on the findings. The aim of this stage is to solve the problem of "how to actively and optimally manage the problems identified" by making optimal recommendations and actionable interventions.

4.2. Explicit-to-Implicit Analytics Evolution

As discussed in Section 2, the last four decades have seen the transfer of data analysis on small and simple data, along with hypothesis testing, to data analytics on large and complex data for hypothesis-free knowledge and insight discovery. Today, the significance and innovation of analytics are better recognized than at any previous time. Correspondingly, a critical question to ask is *What is the conceptual map and evolution of data analytics?* Figure 4 shows a high-level conceptual view of the spectrum and evolution of analytical components and tasks in terms of two major dimensions.

—Dimension 1: Levels of visibility, automation, and state-of-the-art capabilities: that is, the level of data and analytics complexity that is visible to users, the level of automated data analytics, and the level of available capability to handle the complexity and support the automation. With the upgrading of analytics, the visibility of data and analytics becomes lower and the level of automated data analytics is lower too. As data complexity increases, the available capability is weakened. The goal of analytics is to increase the visibility, automation, and capability levels of data understanding, production, and application.

—Dimension 2: Degree of X-complexities, X-intelligence, and value: that is, the degree of data complexity and X-intelligence involved in data and analytics are increased

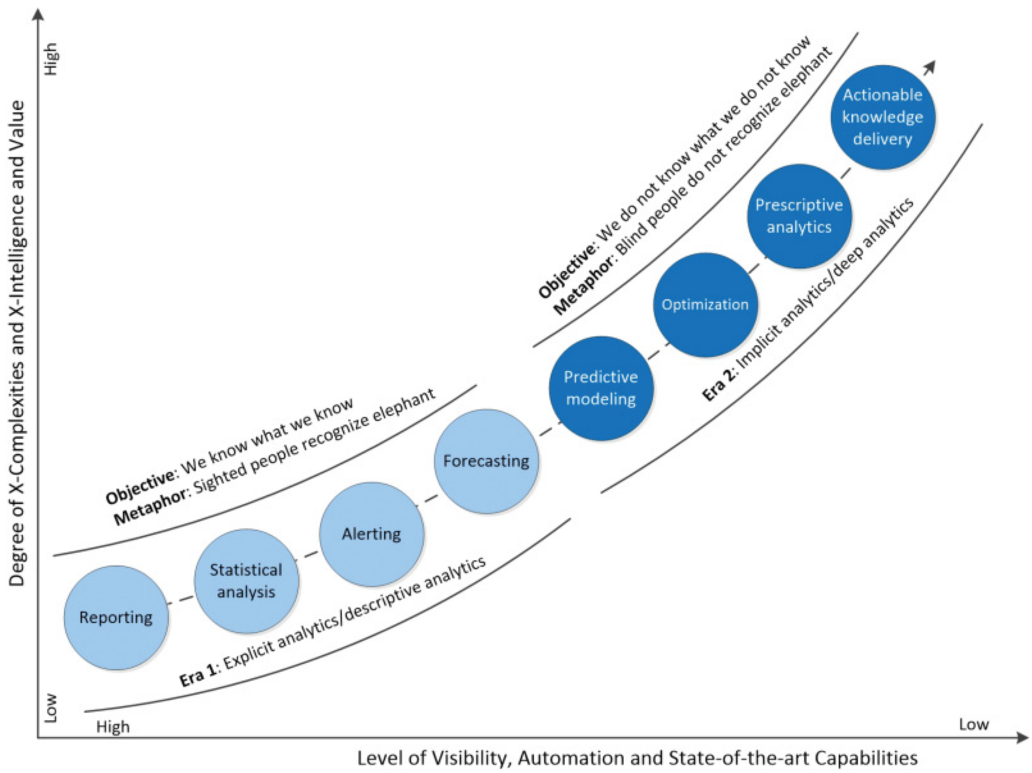

Fig. 4. Explicit-to-implicit analytics spectrum and evolution.

with the movement from lower-level analytics to higher-level analytics. During this process, the level of learned intelligence and value resulting from the corresponding analytics is increased.

As shown in Figure 4, there are many typical analytical approaches and components that may be involved in executing analytics tasks. They include reporting, statistical analysis, alerting, forecasting, predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery (delivering insights-based actions for business decision-making and operations) [Cao et al. 2010]. The listed approaches may be used for nonanalytical purposes, and the corresponding analytical tasks may be addressed by nonanalytical approaches. An example is optimization, which may be used for analytics to select the best options as an analytics approach or may be achieved by findings from analytics approaches as an analytics objective. There may be different foci and connections between the listed analytics approaches. For example, forecasting may be used as an approach for prediction when it focuses on probabilistic estimates of possible futures, while prediction may involve broad techniques and objectives for estimating outcomes.

We also roughly categorize the multiple components and tasks in analytics evolution into two main eras from the perspective of the disciplinary development of analytics:

—Era 1: The era of explicit analytics: which focuses on descriptive analytics. Typical analytics approaches consist of reporting, statistical analysis, alerting, and forecasting.

—Era 2: The era of implicit analytics: which focuses on deep analytics. Typical analytics approaches are predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery.

Note that Figure 4 only shows an evolution path of the analytics family. It does not indicate the path of analytics within a specific organization that conducts and utilizes analytics. It does not indicate a linear path of analytics evolution either. Often, a back-and-forth iterative approach is taken in an analytics team, and multiple analytics components may be involved in parallel for exploring multifaceted observations and understandings.

4.2.1. The Era of Explicit Analytics: Descriptive Analytics. Typical elements and tasks have in the past focused on explicit descriptive analysis and have the following features:

—*Goal*: *we know what we know*, and therefore aim to identify and describe the distribution, generation, and trends of data and business problems;
—*Nature of problem*: similar to *sighted people recognize an elephant*, we know what is to be analyzed by hypothesis-based approaches, and for what purposes;
—*Approach*: domain-driven analysis for which hypotheses are available from domain-specific knowledge and experts; data analysis tests such hypotheses, and the data verifies and explains the hypotheses;
—*Outcome*: focused methods are available from mathematics and statistics as well as from computing. Such methods describe and present what has happened, is happening, or will happen in usually small or highly manipulated data.

4.2.2. The Era of Implicit Analytics: Deep Analytics. The limitation of explicit analytics has recently been more widely recognized in the analytics community, such as in handling latent, uncertain, and non-IID data [Cao 2014, 2015a]. As a result, the focus has recently shifted to *implicit analytics*, and towars *deep analytics*. Deep analytics gains an in-depth understanding of why and how things have happened, are happening, or will happen. Such whys and hows cannot be addressed by descriptive analytics and can determine the next best or worst situation, as well as devise optimal intervention strategies.

—*Goal*: *we do not know what we do not know*, and therefore aim to gain a latent but genuine understanding of data and business problems from visible and invisible sources;
—*Nature of problem*: similar to *blind people recognize an elephant*, we do not know what is to be analyzed, or even why and what we can obtain;
—*Approach*: data-driven discovery by which interesting but hidden insights are learned from data; data creates a view invisible to us and explains the unseen reasons or indicators, to complement domain-driven hypotheses and observations;
—*Outcome*: the focus is on gaining an in-depth, intrinsic, and complete understanding of invisible insights, knowledge and wisdom from data, behaviors, and environment about what has happened, is happening, or will happen in data and business.

Table III summarizes the key categories and features of explicit analytics versus implicit analytics.

4.3. Descriptive-to-Predictive-to-Prescriptive Analytics Paradigm Shift

The paradigm shift from data analysis to data science constitutes the so-called "new paradigm" [Nelson 2009; Hey et al. 2009], that is, data-driven discovery. The history of analytics from the spectrum and dynamics perspective spans two main eras of analytics, as shown in Figure 3. Analytics practices have seen a significant paradigm shift across three major stages: (1) Stage 1: descriptive analytics and reporting, (2) Stage 2:

Table III. Explicit-to-IUmplicit Analytics

Categories	Explicit analytics	Implicit analytics
Nature	Sighted people recognize an elephant	Blind people do not recognize an elephant
Goal	We know what we know	We do not know what we do not know
Approach	Hypothesis + data	Data + environment (incl. domain)
Outcome	Description of data	In-depth representation of data

predictive analytics and business analytics, and (3) Stage 3: prescriptive analytics and decision making.

We briefly discuss these three stages as follows.

—*Stage 1: Descriptive analytics and business reporting*: the major effort is on explicit analytics, which focuses on descriptive analytics and regular and ad hoc reporting. Limited effort is made on implicit analytics for hidden knowledge discovery, which is mainly achieved by using off-the-shelf tools and built-in algorithms. Business reports (often analytical reports) generated by dashboards and automated processes are the means for carrying findings from analytics to management.

—*Stage 2: Predictive analytics and business analytics*: the major effort is on implicit analytics, which focuses on predictive modeling and business analytics (here *business analytics* refers to an in-depth understanding of business through deep analytics. Note that this meaning differs from the broad meaning widely adopted in business and management), with more effort being made to apply forecasting, data mining, and machine learning tools for business understanding and prediction. Patterns, scoring, and findings are presented through dashboards and analytical reports to management.

—*Stage 3: Prescriptive analytics and decision making*: the major effort is on the delivery of recommended optimal (next best) actions for business decisions by discovering invisible knowledge and actionable insights from complex data, behavior, and environment. This is achieved by developing innovative and effective customized algorithms and tools to deeply and genuinely understand domain-specific data and business. Consequently, prescriptive decision-taking strategies, business rules, actions, and recommendations are disseminated to decision-makers for the purpose of taking corresponding actions. By contrast, relatively limited effort is made on explicit analytics since they are conducted through automated processes and systems.

During the paradigm shift (as shown in Figure 5), a significant decrease is seen in the effort made in routine explicit analytics, which is increasingly undertaken by automated analytics services. By contrast, a significant increase in effort is seen in implicit analytics and actionable knowledge delivery [Cao et al. 2010]. The shift from a lower stage to a higher stage accommodates an increasingly higher degree of knowledge, intelligence, and value to an organization.

5. DATA INNOVATION: CHALLENGES AND OPPORTUNITIES

In this section, we summarize the major challenges and opportunities relating to data science in the relevant communities.

There are two ways of exploring major research challenges: one is to summarize the concerns of the relevant communities, and the other is to scrutinize the issues from the perspective of the intrinsic complexity and nature of data science problems as complex systems [Cao and Dai 2008; Cao 2015b]. The first approach summarizes the main topics and issues identified in the statistics [Chambers 1993; Wu 1997; van Dyk et al. 2015], informatics, and computing [Rudin et al. 2014; Cao 2016b] communities, vendors [Stonebraker et al. 2013], government initiatives [USNSF 2012; UN 2010; Government

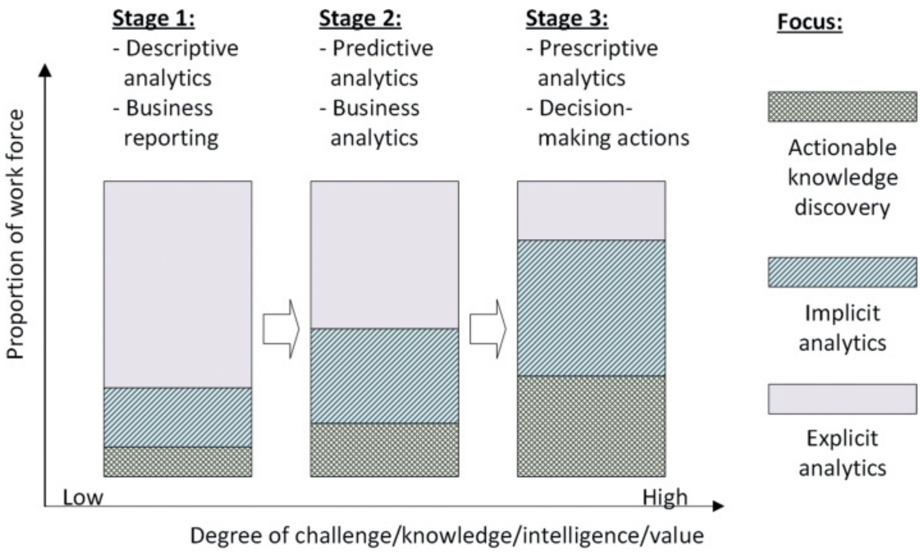

Fig. 5. Descriptive-to-predictive-to-prescriptive analytics paradigm shift.

2015; Commission 2014; UK 2016], and research institutions [UTSAAI 2011; NCSU 2007a] that focus on data science and analytics. This can result in a picture of the main research challenges. The second approach is much more challenging. It requires us to explore the nature of complex data science problems, and the unknown space of the complexities and comprehensive intelligence in complex data systems.

Figure 6 presents a comprehensive conceptual map of data science as a complex system. It summarizes some of the main challenges faced by the data science community in addressing big data complexities [Cao 2016b]. We categorize the challenges facing domain-specific data applications and problems in terms of five major areas:

—*Challenges in data / business understanding*: The challenges here are to identify, specify, represent, and quantify comprehensive complexities, known as X-complexities [Cao 2015b, 2016b]) and intelligence, known as X-intelligence [Cao 2015b, 2016b]). Such X-complexities and X-intelligence cannot be managed well by existing theories and techniques. However, they nonetheless exist and are embedded in domain-specific data and business problems. The issue is to understand in what form, at what level, and to what extent they exist, and to understand how the respective complexities and intelligence interact and integrate with one another. An in-depth understanding of X-complexities and X-intelligence would subsequently result in devising effective methodologies and technologies for incorporating them into data science tasks and processes.
—*Challenges in mathematical and statistical foundations*: The challenges here are to discover and explore whether, how, and why existing theoretical foundations are insufficient, missing, and problematic in disclosing, describing, representing, and capturing the preceding complexities and intelligence and obtaining actionable insights.
—*Challenges in X-analytics and data / knowledge engineering*: The challenge is to develop domain-specific analytic theories, tools, and systems that are not yet available in the body of knowledge. They will represent, discover, implement, and manage the relevant and resultant data, knowledge, and intelligence, and support the engineering of big data storage and management, behavior, and event processing.

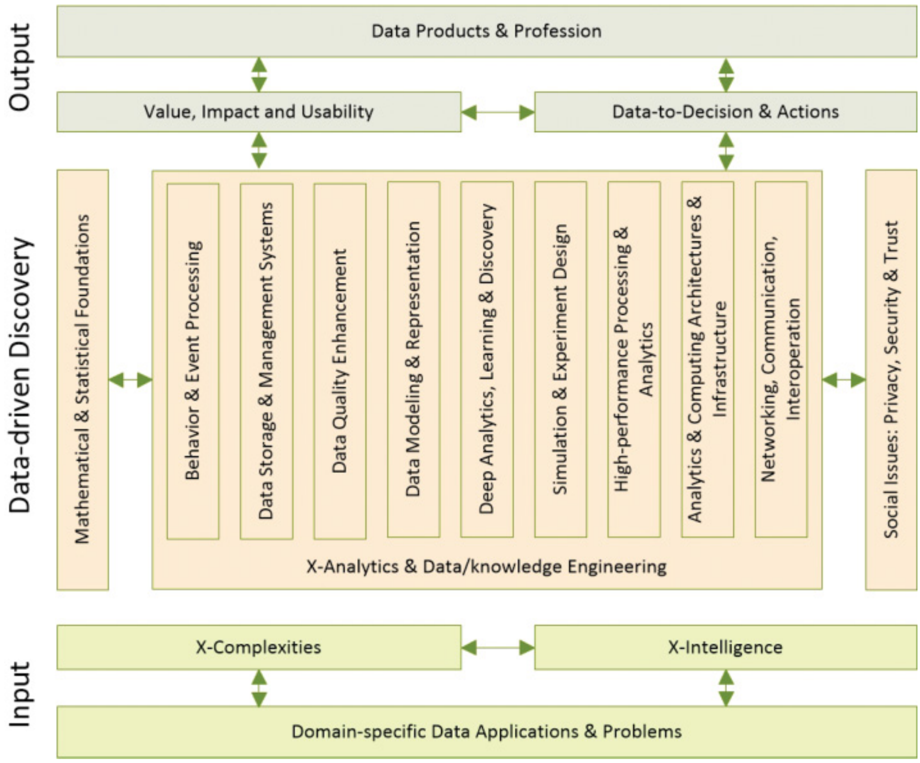

Fig. 6. Data science conceptual map.

—*Challenges in social issues*: This challenge is to identify, specify, and respect social issues related to the domain-specific data and business understanding and data science processes, including processing and protecting privacy, security, and trust and enabling social issues-based data science tasks, which have not, so far, been handled well.
—*Challenges in data value, impact, and usability*: This challenge is to identify, specify, quantify, and evaluate the value, impact, utility, and usability of domain-specific data that cannot be addressed by existing theories and systems, from technical, business, subjective, and objective perspectives.
—*Challenges in data-to-decision and actions*: The challenge recognized here is the need to develop decision-support theories and systems that will enable data-driven decision generation, insight-to-decision transformation, as well as decision-making action generation, and data-driven decision management and governance. These cannot be managed by existing technologies.

The challenges in X-analytics and data/knowledge engineering involve many specific research issues that have not been properly addressed; for example:

—*Behavior and event processing*: how to capture, store, model, match, query, visualize, and manage behaviors and events and their properties, behavior sequences/streams, and the impact and evolution of behaviors and events of individuals and groups in the physical world.
—*Data storage and management systems*: how to design effective and efficient storage and management systems that can handle big data with high volume, velocity, and

variety, and support real-time, online, and on-the-fly processing and analytics; and how to house such data in an Internet-based (including cloud) environment.

—*Data quality enhancement*: how to handle both existing data quality issues, such as noise, uncertainty, missing values, and imbalance that may be present at very different levels due to the significantly increased scale, extent, and complexity of data. At the same time, how to handle new data quality issues emerging in the big data and Internet-based data/business environment, such as cross-organizational, cross-media, cross-cultural, and cross-economic mechanism data science problems.

—*Data modeling, learning, and mining*: how to model, learn, analyze, and mine data that is embedded with comprehensive complexity and intelligence.

—*Deep analytics, learning, and discovery*: how to discover unknown knowledge and intelligence hidden in the space D in Figure 1 (unknown complexities, knowledge, and intelligence; see Can and Fayyad [Cao 2016b]) through inventing new theories and algorithms for implicit and deep analytics that cannot be handled by existing latent learning and descriptive and predictive analytics. Also, how to integrate data-driven and model-based problem-solving that balances common learning models/frameworks and domain-specific data complexity and intelligence-driven evidence learning.

—*Simulation and experimental design*: how to simulate the complexity and intelligence, working mechanisms, processes, dynamics, and evolution in data and business, and how to design experiments and explore the subsequent impact if certain data-driven decisions and actions are undertaken in a business.

—*High-performance processing and analytics*: how to support large-scale, real-time, online, high-frequency, Internet-based (including cloud-based) cross-organizational data processing and analytics while balancing local and global resource involvement and objectives. This requires new batch, array, memory, disk storage, and processing technologies and systems, and massive parallel processing and distributed/parallel and high-performance processing infrastructure, as well as cloud-based processing and storage. It also requires large and complex matrix calculation, mixed data structures and management systems, and data-to-knowledge management.

—*Analytics and computing architectures and infrastructure*: how to facilitate the preceding tasks and processes by inventing efficient analytics and computing architectures and infrastructure based on memory, disk, cloud, and Internet-based resources and facilities.

—*Networking, communication, and interoperation*: how to support the networking, communication, and interoperation between different data science roles in a distributed data science team and during the whole-of-cycle of data science problem-solving. This requires the distributed cooperative management of projects, data, goals, tasks, models, outcomes, workflows, task scheduling, version control, reporting, and governance.

Systematic and interdisciplinary approaches and methodologies are required to address the preceding issues in data science and analytics. These may involve developing a synergy of several research disciplines and areas, including data representation, preprocessing and preparation, information processing, parallel processing, distributed systems, high-performance computing, data management, data warehousing, cloud computing, evolutionary computation, neural networks, fuzzy systems, enterprise infrastructure, system architecture, communication and networking, integration and interoperation, machine learning, data modeling, analytics and mining, service computing, system simulation, experimental design, and evaluation. It may also involve business and social aspects, including industry transformation, enterprise information systems, business intelligence, business process management, project management, information security, trust and reputation, privacy processing, business impact modeling, business value, and utility evaluation [DSAA 2014; Cao 2016a]. Interdisciplinary initiatives are necessary to bridge the gaps between the respective disciplines, and to create new opportunities for the invention and development of new technologies, theories, and tools to address critical complexities in complex data science problems that cannot be addressed by singular disciplinary efforts.

6. DATA ECONOMY: DATA INDUSTRIALIZATION AND SERVICES

Data science and big data analytics have led to next-generation economy innovation, competition, and productivity [McKinsey 2011], as shown by the rapidly updated Big Data Landscape [BDL 2016b]. Significant new business opportunities and previously impossible prospects have become possible through the creation of data products, data economy, and data industrialization and services [Yiu 2012; CSC 2012; IBM 2016a; Loukides 2012]. In this section, we discuss such opportunities.

6.1. Data Industry

If data is viewed in the same way as oil, as the new international currency, then clearly the global economy is experiencing a revolutionary change from data poor to data rich and data-driven. On one hand, data industrialization creates new business, where companies, organizations, and even countries compete over how to best use data to create new data products. On the other hand, core businesses, including retail business and manufacturing, are giving way to a new economy that is centered on the data industry and digital economy. This is evidenced by the domination of data-enabled companies listed in the top 10 global companies, especially the largest data company, Google, and the largest Initial Public Offering Alibaba.

The data industry is taking shape and gaining significance as a driving force in the new global economy. Without loss of generality, Figure 7 illustrates those aspects in which new data business and the resultant areas of data business may grow. The main driving forces of the data industry come from the following six core areas: data/analytics design, data/analytics content, data/analytics software, data/analytics infrastructure, data/analytics services, and data/analytics education.

—*Data/analytics design* includes the invention of new methods and ways of designing and producing digital and data products, services, business models, engagement models, communication models, pricing modeling, economic forms, value-added data products/services, decision support systems, automation systems, and tools;
—*Data/analytics content* includes acquiring, producing, maintaining, publicizing, disseminating, recommending, and presenting data-centered content through online, mobile, social media platforms, and other channels;
—*Data/analytics software* refers to the creation of software, platforms, architectures, services, tools, systems, and applications that acquire, organize, manage, analyze, visualize, use, and present data for specific business and scientific purposes, and provide quality assurance to support these aspects;
—*Data/analytics infrastructure* relates to creating infrastructure and devices for data storage, backup, server revenue, data centers, data management and storage, cloud, distributed, and parallel computing infrastructure, high-performance computing infrastructure, networking, communications, and security;
—*Data/analytics services* focus on providing strategic and tactical thinking leadership, technical and practical consulting services, problem-oriented solutions and applications, outsourcing, and specific services for data auditing and quality enhancement, data collection, extraction, transformation and loading, recommendation, data center/infrastructure hosting, data analytics, and more;

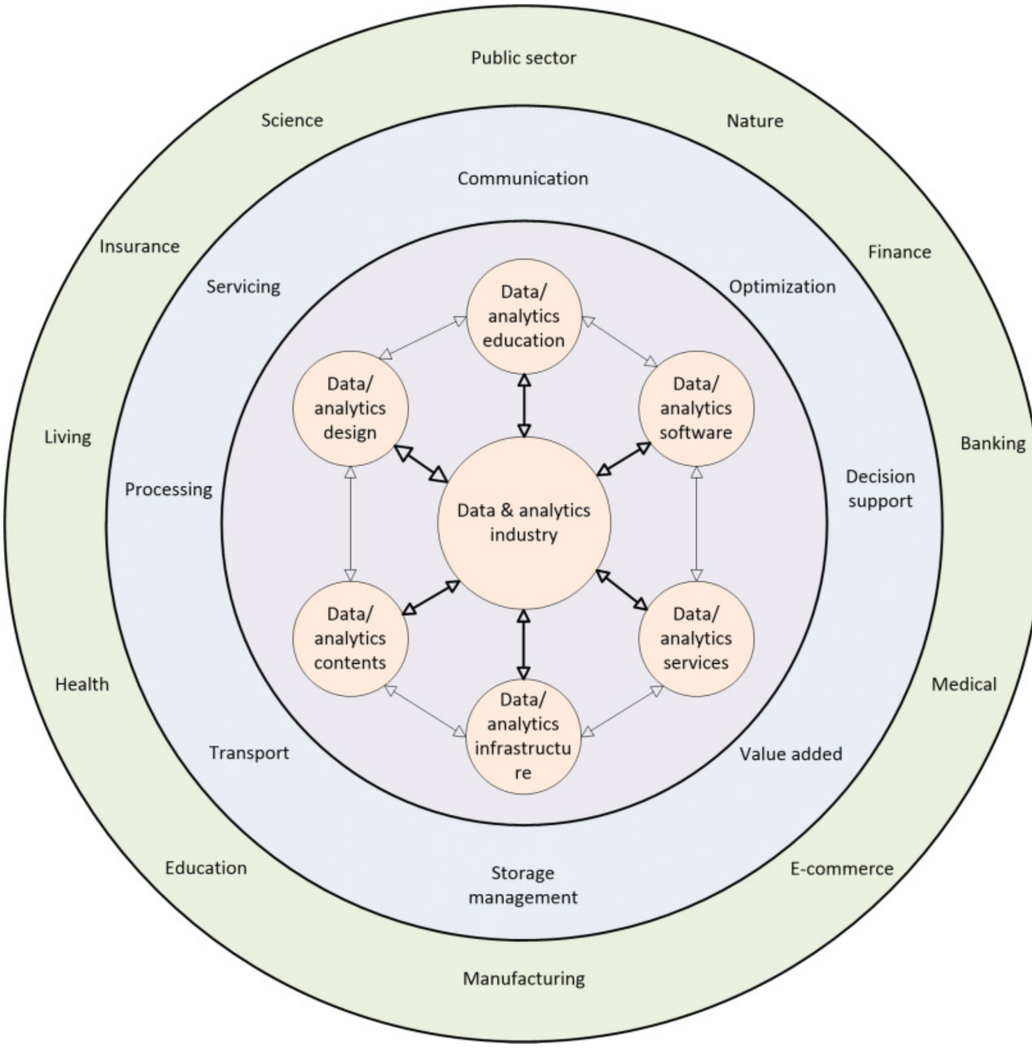

Fig. 7. Data and analytics-enabled industry and business transformation.

—*Data / analytics education* enables the building of corporate competency and training, as well as offering online/offline/degree-based courses, workshops, materials, and services that will allow the gaps in the supply of qualified data professionals to be filled, thus contributing to building and enhancing the community of this discipline.

The preceding six core data/analytics areas will see the growth of new data business in terms of the following core aspects and procedures: data storage and management, understanding, processing, optimization, value-added opportunities, transport and communication, servicing, and decision support.

The respective core data/analytics sectors and core procedures can be developed in any data-related sectors, especially data-intensive domains and sectors such as telecommunication, government, finance, banking, capital markets, lifestyle, and education. Core business including manufacturing and living business will see increased opportunities for better collection, management, and use of data. Analytics will be undertaken to improve productivity, effectiveness, and efficiency, and create new value-added growth in the economy.

Interestingly, as we have seen, the data economy occupies the top 10 largest capital entities. The data industry continues to create new business models, products, services, operationalization modes, and workforce models. Data economy will further change the way we live, work, learn, and are entertained, as new facilities and environments are created in which data plays a critical role.

6.2. Data Services

Data services form part of the whole landscape of data and analytics which, as noted previously, is changing every aspect of the way we live. Data services can be differentiated from traditional services by the fact that they are not traditional physical material- or energy-oriented services.

—Data services act as the core business rather than the auxiliary business of an economy.
—Data-driven production and decision-making emerges as the core function in large organizations for complex decision-making and strategic planning, rather than adjunct facilities.
—Data services are online, mobile ,and socially based, embedded in our activities and agenda.
—Data business is global and 24/7, offered at any place at any time on demand or in a supply-driven mode.
—The provision of data services does not require traditional production elements such as intensive workshops, factories, and office facilities.
—Data-driven services offer real-time public service data management, high-performance processing, analytics, and decision-making.
—Data-driven services support full lifecycle analysis, from descriptive, predictive, and prescriptive analytics for the prediction, detection, and prevention of risk, to innovation and optimization.
—Data-analytical services are intelligent or can enhance the intelligence of generous data and information services.
—Data services enable cross-media, cross-source, and cross-organization innovation and practice.
—Data services demonstrate significant savings and efficiency improvement through the delivery of actionable knowledge/insights.

Some typical data services delivered through analytics for both core business and new economy are listed next as examples:

—*Credit scoring*: to establish the credit worthiness of a customer requesting a loan.
—*Fraud detection*: to identify fraudulent transactions and suspicious behavior.
—*Healthcare*: to detect overservice, underservice, fraud, and events like epidemics;
—*Insurance*: to detect fraudulent claims and assess risk.
—*Manufacturing process analysis*: to identify the causes of manufacturing problems and to optimize the processes.
—*Marketing and sales*: to identify potential customers and establish the effectiveness of campaigns.
—*Portfolio trading*: to optimize a portfolio of financial instruments by maximizing returns and minimizing risk.
—*Surveillance*: to detect intrusion, objects, persons, and linkages from multisensor data and remote sensing.

—*Understanding customer behaviors*: to model churn, affinities, propensities, and next best actions on intervention behaviors.
—*Web analytics*: to model user preferences from data to devise and provide personalized and targeted services.

A major challenge and increasing need in the data industry is to provide global or Internet-based data services for a collection of organizations, such as multinational companies and whole-of-government. Such services need to

—define global data analytics objectives and benefits;
—support good data governance, security, privacy, and accountability to enable smarter data use and sharing;
—support data matching and sharing in the context of cross-organizational, cross-platform, cross-format, and cross-analytical goals;
—prepare global and organization-specific local/departmental data;
—foster global and local analytics capabilities, capacity, and competency;
—enable sharing and collaboration in data and analytics skills, infrastructure, tools, techniques, and outcomes;
—support crowdsourcing, collaborative and parallel analytic tasks, and analytic workflow management;
—support analytic capability and package sharing;
—support data and data software versioning management and control [Bhardwaj et al. 2015] at a global and collaborative level; and
—support the visualization and dissemination of outcomes to targeted audiences and in personalized preferences.

Data-driven industry and service are forming new trends in data science for business; for instance:

—Advanced analytics is no longer just for analysts [Ghodke 2015]; dummy analytics is becoming the default setting of management and operational systems.
—Cloud data management, storage, and cloud-based analytics are gaining popularity [Ghodke 2015] and are replacing traditional management information systems, business support systems, and operational support systems.
—Data science on scale from multiple sources of data is becoming feasible; Internet-based services are a strongly growing area of the new economy.
—Analytics as a service is becoming feasible with appropriate social issue management, as analytics becomes a reality everywhere and is embedded in business, mobile, social, and online services.
—Visual analytics is becoming a common language.
—Data services can be mixed with virtual reality and presented in a way that combines physical and virtual worlds, resources, and intelligence.
—Services on matched and mixed data are streamlined into a one-world process, with both local and global objectives addressed.

7. DATA EDUCATION: CAPABILITIES AND COMPETENCY

Data innovation and economy is dependent on the corresponding data and analytics capabilities and competencies and the ability to handle related social issues. These are weak areas, and there are significant gaps in the current body of knowledge, organizational maturity [Paulk et al. 1993; Crowston and Qin 2011], education, and training. The requirement is to "think with data," "manage data," "compute with data," "mine on data," "communicate with data," "deliver with data," and "take action on data" [Cao 2016c]. This section discusses these important matters.

More and more industry and government organizations recognize the value of data for decision-making and have set up general and specific data scientist roles to support data science and engineering, for example, Chief Data Officer, Chief Analytics Officer, data modelers and data miners, in addition to data engineers and business analysts.

7.1. Data Scientists in a Sexy Profession

The role of the data scientist was recognized 10 years ago, and it has become a sexy profession in the job market. The next-generation data scientists will be mostly welcomed in the increasingly important data economy and data-to-decision.

In 2004, Dr. Usama Fayyad was appointed as the Chief Data Officer of Yahoo, which opened the door to a new career possibility: the *data science professional* [Manieri et al. 2015; Harris et al. 2013] or more specifically, *data scientist*, for those people whose role very much centers on data. In 2015, the White House appointed the first U.S. Chief Data Scientist [Whitehouse 2015]. This role "will shape policies and practices to help the U.S. remain a leader in technology and innovation, foster partnerships to help responsibly maximize the nation's return on its investment in data, and help to recruit and retain the best minds in data science to join us in serving the public." [Whitehouse 2015].

Today, the role of data scientist [Patil 2011] is regarded as "the sexiest job of the 21st century" [Davenport and Patil 2012]. It is reported that data scientists earn much higher salaries than those in other data-related jobs, with a median salary of US$120k for data scientists and US$160k for managers, according to the 2014 Burtchworks survey [Burtch 2014]. This is attributed to the fact that 88% of respondents in this survey have at least a Master's degree, while 46% also hold a Doctorate compared to only 20% of other Big Data professionals. In the 2015 O'Reilly survey [King and Magoulas 2015], 23% were found to hold a doctorate, while another 44% had a Master's. The median annual base salary of this survey sample was US$91,000 globally, and among US respondents was US$104,000, compared to US$150k for "upper management" (higher than project and product managers).

7.2. What Does a Data Scientist Do

So, what are the roles and responsibilities of data scientists? Here we summarize the findings from several documents on government initiatives:

—The US National Science Board defines data scientists as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection." [NSB 2005]
—In a report from the US Committee on Science of the National Science and Technology Council, data scientists are defined as "Scientists who come from information or computer science backgrounds but learn a subject area and may become scientific data curators in disciplines and advance the art of data science. Focus on all parts of the data life cycle." [CSNSTC 2009]
—The Joint Information Systems Committee defines data scientists as "people who work where the research is carried out, or, in the case of data centre personnel, in close collaboration with the creators of the data and may be involved in creative inquiry and analysis, enabling others to work with digital data, and developments in data base technology." [Swan and Brown 2008]

In business, immense interest has been expressed by multinational vendors, social media and online communities, and information providers, such as IBM [2016b], LinkedIn [2016], Kdnuggets [2016], Facebook [2016], and SIAM [2016] about the roles and responsibilities of data scientists and what makes a good data scientist.

For instance, a data scientist metromap was created in Chandrasekaran [2013]. The metromap covers 10 areas and domains: Fundamentals, Statistics, Programming, Machine Learning, Text Mining/Natural Language Processing, Data Visualization, Big Data, Data Ingestion, Data Munging, and Toolbox. Each area/domain is represented as a "metro line," with the stations depicting the topics to be learned and understood in a progressive fashion. In addition, INFORMS summarizes the following seven job tasks for data scientists [INFORMS 2014]: Business Problem (Question) Framing, Analytics Problem Framing, Data, Methodology (Approach) Selection, Model Building, Deployment, and Model Lifecycle Management.

An increasing number of academic and research institutions are working on defining the certification and accreditation of next-generation data scientists. This is reflected in general and domain-specific data science curricula for Masters and PhD qualifications, such as a PhD in Analytics [UTS 2011] and a Master's degree in SCM predictive analytics [INFORMS 2014].

Without the loss of generality, typical domain-free and problem-neutral responsibilities and requirements for jobs announced in social media channels (e.g., Google Groups, Facebook, and LinkedIn) and what we have experienced in the past 15 years in large governmental and business organizations can be summarized as follows:

—Learn the business problem domain, talk to business experts and decision-makers to understand the business objectives, requirements and preferences, issues and constraints facing an organization; understand the organizational maturity; identify, specify, and define the problems, boundaries, and environment, as well as the challenges; generate business understanding reports.
—Identify and specify social and ethical issues such as privacy and security; develop ethical reasoning plans to address social and ethical issues.
—Understand data characteristics and complexities; identify the problems and constraints of the data; develop a data understanding report; specify and scope analytical goals and milestones by developing respective project plans to set up an agenda and create governance and management plans.
—Set up engineering and analytical processes corresponding to analytical goals for turning business and data into information, turning information into insight, and turning insight into business decision-making actions by developing technical plans for the discovery, upgrade, and deployment of relevant data intelligence.
—Transform business problems into analytical tasks, and conduct advanced analytics by developing corresponding techniques, models, methods, algorithms, tools and systems, experimental design and evaluation of data science, generating better practices experience, performing descriptive, predictive, and prescriptive analytics, conducting survey research, and supporting visualization and presentation.
—Based on the understanding of data characteristics and complexities, extract, analyze, construct, mine, and select discriminative features, constantly optimize and innovate new variables for best possible problem representation and modeling; when necessary, conduct data quality enhancement [Hazena et al. 2014].
—Combine analytical, statistical, algorithmic, engineering, and technical skills to mine relevant data by involving contextual information; invent novel and effective models, and constantly improve modeling techniques to optimize and boost model performance and seek to achieve best practice.
—Maintain, manage, and refine projects and milestones, and their processes, deliverables, evaluation, risk, and reporting to build active, lifecycle management.
—Develop corresponding services, solutions, and products or modules to feed into a system package on top of user-specified programming languages, frameworks, and infrastructure, or open source tools and frameworks.

—Maintain the privacy, security, and veracity of data and deliverables.
—Engage in frequent client interaction during the whole lifecycle; tell clear and concise stories and draw simple conclusions from complex data or algorithms; provide clients with situational analyses and deep insights into areas requiring improvement; translate into business improving actions in the final deployment.
—Write coherent reports and make presentations to specialists and nonspecialists; present executive summaries with precise and evidence-based recommendations and risk management strategies, especially for decision-makers and business owners.

7.3. What Makes a Good Data Scientist

To satisfy the preceding position requirements, data scientist candidates need to have certain qualifications in addition to the analytic skills that are the foundation of this role. These qualifications and abilities include the following:

—Thinking, mindset, and ability to think analytically, creatively, critically, and inquisitively.
—Methodologies and knowledge of complex systems and approaches for conducting both top-down and bottom-up problem-solving.
—Master's or PhD degree in computer science, statistics, mathematics, analytics, data science, informatics, engineering, physics, operations research, pattern recognition, artificial intelligence, visualization, information retrieval, or related fields.
—A deep understanding of common statistics, data mining, and machine learning methodologies and models.
—Ability to implement, maintain, and troubleshoot big data infrastructure, such as cloud computing, high-performance computing infrastructure, distributed processing paradigms, stream processing, and databases.
—Knowledge of human-computer interactions, visualization, and knowledge representation and management;
—Background in software engineering (including systems design and analysis), quality assurance.
—Experience working with large datasets, and mixed data types and sources in a networked and distributed environment.
—Experience in data extraction and processing, feature understanding and relation analysis.
—Active interest and knowledge in multidisciplinary and trans-disciplinary studies and methods in scientific, technical, and social and life sciences.
—Substantial experience with state-of-the-art analytics-oriented scripting, data structures, programming languages, and development platforms in a Linux, cloud, or distributed environment.
—Theoretical background and domain knowledge for the evaluation of the technical and business merits of analytic findings.
—Excellent written and verbal communication [Matsudaira 2015] and organizational skills, ability to write and edit analytical materials and reports for different audiences, and capacity to transform analytical concepts and outcomes into business-friendly interpretations; ability to communicate actionable insights to nontechnical audiences, and experience in data-driven decision making.

While there is significant role overlap between *data scientists* and *Business Intelligence (BI) professionals* [SAS 2013], different research works show that data science professionals are generally much more data and technology-savvy rather than business-oriented, with most holding a Master's or PhD degree in statistics or computer science.

An EMC data science community survey [EMC 2011] shows that (1) data scientists can open up new possibilities; (2) compared to 37% of BI professionals trained in business, 24% of data science professionals are in computer science, 17% are in engineering, and 11% are in hard science; (3) compared to BI toolkits, data science toolkits are more technically sophisticated and more diversified; (4) the number of data scientists undertaking data experiments is almost double that of BI professionals; (5) data science professionals more frequently interact with diverse technical and business roles in an organization (such as data scientists, strategic planners, statisticians, marketing staff, sales people, graphic designers, business management and IT administration, programmers, and HR personnel) than BI professionals; (6) compared to working on normal data, big data manipulators tend to tackle more sophisticated data complexities; and (7) data science professionals spend almost double the time they spend on normal data on big data manipulation (e.g., data parsing, organization, mining, algorithms, visualization, story-telling, dynamics, and decisions).

As a data-centric expert, a good data scientist is also expected to know the underlying domain well. Without an in-depth understanding of the domain, the actionability of the data deliverables and products by data scientists may be low. However, a data scientist is no substitute for domain experts in complex data science problem-solving [Cao 2016b]. Similar to any other disciplinary specialists, data scientists work more effectively by collaborating with domain-specific specialists and subject matter experts to achieve broader impact. This is similar to the requirements of domain-driven, actionable knowledge discovery [Cao et al. 2010].

7.4. Tools for Data Scientists

In the previous sections, the respective responsibilities and qualifications of data scientists have been discussed. In Section 5, relevant research challenges and issues have been listed. To support the data services listed in Section 6.2, the corresponding roles and qualifications discussed in Sections 7.2 and 7.3 are necessary.

In this section, we discuss tools that may be used by data scientists to address the preceding aspects. Tools are categorized in terms of cloud infrastructure, data and application integration, data preparation and processing, analytics, visualization, programming, master data management, high-performance processing, business intelligence reporting, and project management. A data scientist may use one or more of these tools on demand for data science problem-solving.

—Cloud infrastructure: Such as Apache Hadoop, Spark, Cloudera, Amazon Web Services, Unix shell/awk/gawk, 1010data, Hortonworks, Pivotal, and MapR. Most traditional IT vendors have migrated their services and platforms to support cloud.
—Data/application integration: Including Ab Initio, Informatica, IBM InfoSphere DataStage, Oracle Data Integrator, SAP Data Integrator, Apatar, CloverETL, Information Builders, Jitterbit, Adeptia Integration Suite, DMExpress Syncsort, Pentaho Data Integration, and Talend [Review 2016].
—Master data management: Typical software and platforms include IBM InfoSphere Master Data Management Server, Informatica MDM, Microsoft Master Data Services, Oracle Master Data Management Suite, SAPNetWeaver Master Data Management tool, Teradata Warehousing, TIBCO MDM, Talend MDM, Black Watch Data.
—Data preparation and processing: In Today [2016], 29 data preparation tools and platforms were listed, such as Platfora, Paxata, Teradata Loom, IBM SPSS, Informatica Rev, Omniscope, Alpine Chorus, Knime, and Wrangler Enterprise and Wrangler.
—Analytics: In addition to well-recognized commercial tools including SAS Enterprise Miner, IBM SPSS Modeler and SPSS Statistics, MatLab, and RapidMiner [2016],

many new tools have been created, such as DataRobot [2016], BigML [2016], MLBase [Lab 2016], and APIs including Google Cloud Prediction API [Google 2016b].
—Visualization: Many free and commercial software are listed in KDnuggets [2015] for visualization, such as Interactive Data Language, IRIS Explorer, Miner3D, NETMAP, Panopticon, ScienceGL, Quadrigram, and VisuMap.
—Programming: In addition to the main languages R, SAS, SQL, Python, and Java, many others are used for analytics, including Scala, JavaScript, .net, NodeJS, Obj-C, PHP, Ruby, and Go [Davis 2016].
—High-performance processing: In Wikipedia [2016a], about 40 computer cluster software are listed and compared in terms of their technical performance, such as Stacki, Kubernetes, Moab Cluster Suite, and Platform Cluster Manager.
—Business intelligence reporting: There are many reporting tools available [Capterra 2016b; Wikipedia 2016c], typical of which are Excel, IBM Cognos, MicroStrategy, SAS Business Intelligence, and SAP Crystal Reports.
—Project management: In Capterra [2016a], more than 500 software and tools were listed for project management, including Microsoft Project, Atlassian, Podio, Wrike, Basecamp, and Teamwork.
—Social network analysis: In Desale [2015], 30 tools were listed for SNA and visualization, such as Centrifuge, Commetrix, Cuttlefish, Cytoscape, EgoNet, InFlow, JUNG, Keynetiq, NetMiner, Network Workbench, NodeXL, and SocNetV (Social Networks Visualizer).
—Other tools: Increasing numbers of tools have been developed and are under development for domain-specific and problem-specific data science, such as Alteryx and Tableau for tablets; SuggestGrid and Mortar Recommendation Engine for recommender systems [Github 2016b]; OptumHealth, Verisk Analytics, MedeAnalytics, McKesson and Truven Health Analytics [Technavio 2016] for healthcare analytics; BLAST, EMBOSS, Staden, THREADER, PHD, and RasMol for bioinformatics.

8. THE FUTURE OF DATA SCIENCE

There is continuing debate about how data science will evolve in the next 50 years and what it will ultimately look like. With the joint efforts to be made by the entire scientific community, data science will build its systematic scientific foundations, disciplinary structure, theoretical systems, technological families, and engineering tool sets as an independent science.

The last 50 years since the proposal of the concept "data science" has contributed to the progressive and now widespread acceptance of the need for a new science and its initial conceptualization through its transition and transformation from statistics to the merger with existing disciplines and fields. The next 50 years of data science will extend beyond statistics to identify, discover, explore, and define specific foundational scientific problems and grand challenges. It will build a systematic family of scientific methodologies and methods and self-contained disciplinary systems and curricula that are not merely a relabeled "salad" created by mixing existing disciplinary components.

Based on the understanding of the intrinsic challenges and nature of data science [Cao 2016b, 2016c], the development of data science may seek to:

—Design and develop *data brain* that can autonomously mimic human brain working mechanisms to recognize, understand, analyze, and learn data and environment, infer and reason about knowledge and insight, and correspondingly decide actions.
—Deepen our *understanding of data invisibility* (i.e., *invisible data characteristics, complexities, intelligence, and value*), in particular, to understand their X-complexities and X-intelligence (see Cao and Fayyad [Cao 2016b]). The exploration of what we

do not know about what we do not know will strengthen our understanding of the capabilities, limitations, and future directions of data science.
—Broaden conceptual, theoretical, and technological systems for data science by enabling cross-disciplinary and trans-disciplinary research, innovation, and education. This will address existing issues such as the variations in statistics hypotheses and will discover and propose problems that are currently invisible to broad science or specific fields.
—Invent *new data representation capabilities*, including designs, structures, schemas and algorithms to make invisible data complexities and unknown characteristics in complex data more visible and explicit, and more easily understood or explored.
—Design *new storage, access, and management mechanisms*, including memory, disk, and cloud-based mechanisms, to enable the acquisition, storage, access, sampling, and management of richer characteristics and properties in the physical world that have been simplified and filtered by existing systems, and to support scalable, transparent, flexible, interpretable, and personalized data manipulation and analytics in real time.
—Create *new analytical and learning capabilities*, including original mathematical, statistical, and analytical theories, algorithms, and models, to disclose the unknown knowledge in unknown space.
—Build *new intelligent systems and services*, including corporate and Internet-based collaborative platforms and services, to support the automated, or human-data-cooperative, collaborative, and collective exploration of invisible and unknown challenges in unknown space.
—Train *the next-generation data scientists and data professionals* who are qualified for data science problem-solving, with data literacy, thinking, competency, consciousness, curiosity, communication, and cognitive intelligence, to work on the preceding data science agenda.
—Assure cross-domain and trans-disciplinary cooperation, collaborations, and alliance in complex data science problem-solving. This requires the education of competent data scientists who are multidisciplinary experts, as well as collaboration between data scientists and domain-specific experts.
—Discover and invent *data power* as yet unknown to current understanding and imagination, such as new data economy, mobile applications, social applications, and data-driven business.

9. CONCLUSIONS

Data science, big data, and advanced analytics have been increasingly recognized as major driving forces for next-generation innovation, economy, and education. Although they are at an early stage of development, strategic discussions about the big picture, trends, major challenges, future directions, and prospects are critical for the healthy development of the field and the community. The purpose of this article has been to share an overview of the conceptualization, development, observations, and thinking about the age of data science initiatives, research, innovation, industrialization, profession, competency, and education.

We are witnessing a highly evolving data world that seamlessly connects to our daily life, work, learning, economy, and entertainment. New efforts are increasingly being made by government, industry, academia, and even private institutions on ways to convert data for decision-making, and promote the research and development of data science and analytics. The next generation of data science, encompassing a broad range of disciplines, science, and economy, relies heavily on the strategic planning and visionary actions that will be undertaken in prioritized data research areas and start-ups. Without any doubt, today's questions such as "why do we need data science" will be replaced by a family of scientific theories and tools to address the visible grand challenges and significant problems facing tomorrow's big data, science, business, society, and the economy. We will be greatly amazed by the surprising developments and potential changes that will take place in the next 50 years.

REFERENCES

ACEMS. 2014. The Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers. Retrieved from acems.org.au/.

Ritu Agarwal and Vasant Dhar. 2014. Editorial-big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research* 25, 3 (2014), 443–448.

Xinhua News Agency. 2016. The 13th Five-Year Plan for the National Economic and Social Development of the People's Republic of China. Retrieved from http://news.xinhuanet.com/politics/2016lh/2016-03/17/c_1118366322.htm.

AGIMO. 2013. AGIMO Big Data Strategy - Issues Paper. Retrieved from www.finance.gov.au/files/2013/03/Big-Data-Strategy-Issues-Paper1.pdf.

Paul E. Anderson, James F. Bowring, Rene McCauley, George Pothering, and Christopher W. Starr. 2014. An undergraduate degree in data science: Curriculum and a decade of implementation experience. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE'14)*. 145–150.

ASA. 2015. ASA views on data science. Retrieved from http://magazine.amstat.org/?s=data+science&x=0&y=0.

AU. 1990. Data-matching Program. Retrieved from http://www.comlaw.gov.au/Series/C2004A04095.

AU. 2010. Declaration of Open Government. Retrieved from http://agimo.gov.au/2010/07/16/declaration-of-open-government/.

AU. 2013. Attorney-General's Department. Retrieved from http://www.attorneygeneral.gov.au/Mediareleases/Pages/2013/Seconder/22May2013-AustraliajoinsOpenGovernmentPartnership.aspx.

AU. 2016. Australia Big Data. Retrieved from http://www.finance.gov.au/big-data/.

Kayode Ayankoya, André P. Calitz, and Jean Greyling. 2014. Intrinsic relations between data science, big data, business analytics and datafication. *ACM International Conference Proceeding Series* 28 (2014), 192–198.

John Bailer, Roger Hoer, David Madigan, Jill Montaquila, and Tommy Wright. 2012. Report of the ASA workgroup on master's degrees. Retrieved from http://magazine.amstat.org/wp-content/uploads/2013an/masterworkgroup.pdf.

Ben Baumer. 2015. A data science course for undergraduates: Thinking with data. *The American Statistician* 69, 4 (2015), 334–342.

BDL. 2016a. Big Data Landscape. Retrieved from www.bigdatalandscape.com.

BDL. 2016b. Big Data Landscape 2016 (Version 3.0). Retrieved from http://mattturck.com/2016/02/01/big-data-landscape/.

Mark A. Beyer and Douglas Laney. 2012. The Importance of 'Big Data': A Definition. Retrieved from https://www.gartner.com/doc/2057415 Gartner.

Anant Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshp, Aaron J. Elmore, Samuel Madden, and Aditya Parameswaran. 2015. Datahub: Collaborative data science & dataset version management at scale. In *CIDR*.

BigML. 2016. BigML. Retrieved from https://bigml.com/.

Kirk D. Borne, Suzanne Jacoby, Karen Carney, Andy Connolly, Timothy Eastman, M. Jordan Raddick, J. A. Tyson, and John Wallin. 2010. The revolution in astronomy education: Data science for the masses. Retrieved from http://arxiv.org/pdf/0909.3895v1.pdf.

Sebastien Boyer, Ben U. Gelman, Benjamin Schreck, and Kalyan Veeramachaneni. 2015. Data science foundry for MOOCs. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA'15)*. 1–10.

Leo Breiman. 2001. Statistical modeling: The two cultures. *Statistical Science* 16, 3 (2001), 199–231.

Gavin Brown. 2009. Review of Education in Mathematics, Data Science and Quantitative Disciplines: Report to the Group of Eight Universities. Retrieved from https://go8.edu.au/publication/go8-review-education-mathematics-data-scie nce-and-quantitative-disciplines.

Linda Burtch. 2014. The Burtch Works Study: Salaries of Data Scientists. Retrieved from http://www.burtchworks.com/files/2014/07/Burtch-Works-Study_DS_final.pdf.

Kanyarat Bussaban and Phanu Waraporn. 2015. Preparing undergraduate students majoring in computer science and mathematics with data science perspectives and awareness in the age of big data. In *Proceedings of the 7th World Conference on Educational Sciences*, Vol. 197. 1443–1446.

CA. 2016. Canada Capitalizing on Big Data. http://www.sshrc-crsh.gc.ca/news_room-salle_de_presse/latest_news-nouvell es_recentes/big_data_consultation-donnees_massives_consultation-eng.aspx.

Longbing Cao. 2010a. Domain driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering* 22, 6 (2010), 755–769.

Longbing Cao. 2010b. In-depth behavior understanding and use: The behavior informatics approach. *Information Science* 180, 17 (2010), 3067–3085.

Longbing Cao. 2011. Strategic Recommendations on Advanced Data Industry and Services for the Yanhuang Science and Technology Park.

Longbing Cao. 2014. Non-IIDness learning in behavioral and social data. *The Computer Journal* 57, 9 (2014), 1358–1370.

Longbing Cao. 2015a. Coupling learning of complex interactions. *Journal of Information Processing and Management* 51, 2 (2015), 167–186.

Longbing Cao. 2015b. *Metasynthetic Computing and Engineering of Complex Systems*. Springer.

Longbing Cao. 2016a. Data science and analytics: A new era. *International Journal of Data Science and Analytics* 1, 1 (2016), 1–2.

Longbing Cao. 2016b. Data science: Challenges and directions. Technical Report, UTS Advanced Analytics Institute.

Longbing Cao. 2016c. Data Science: Nature and Pitfalls. Technical Report, UTS Advanced Analytics Institute.

Longbing Cao. 2016d. Data Science: Profession and Education. Technical Report, UTS Advanced Analytics Institute.

Longbing Cao. 2017. *Understand Data Science* (to be published). Springer.

Longbing Cao and Ruwei Dai. 2008. *Open Complex Intelligent Systems*. Post Telecom Press.

Longbing Cao, Ruwei Dai, and Mengchu Zhou. 2009. Metasynthesis: M-space, m-interaction and m-computing for open complex giant systems. *IEEE Transactions on Systems, Man, and Cybernetics–Part A* 39, 5 (2009), 1007–1021.

Longbing Cao and Philip S. Yu (Eds). 2012. *Behavior Computing: Modeling, Analysis, Mining and Decision*. Springer.

Longbing Cao, Yuming Ou, and Philip S Yu. 2012. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering* 24, 8 (2012), 1378–1392.

Longbing Cao, Philip S. Yu, Chengqi Zhang, and Yanchang Zhao. 2010. *Domain Driven Data Mining*. Springer.

Capterra. 2016a. Top Project Management Tools. Retrieved from http://www.capterra.com/project-manageme nt-software/.

Capterra. 2016b. Top Reporting Software Products. Retrieved from http://www.capterra.com/reporting-software/.

CBDIO. 2016. China Big Data Industrial Observation. Retrieved from www.cbdio.com.

CCF-BDTF. 2013. China Computer Federation Task Force on Big Data. Retrieved from http://www.bigdataforum.org.cn/.

John M. Chambers. 1993. Greater or lesser statistics: A choice for future research. *Statistics and Computing* 3, 4 (1993), 182–184.

Swami Chandrasekaran. 2013. Becoming a Data Scientist. Retrieved from http://nirvacana.com/thoughts/becoming-a-data-scientist/.

Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36, 4 (2012), 1165–1188.

China Information Security. 2015. Big Data Strategies and Actions in Major Countries. Retrieved from http://www.cac.gov.cn/2015-07/03/c_1115812491.htm.

Thomas R. Clancy, Kathryn H. Bowles, Lillee Gelinas, Ida Androwich, Connie Delaney, Susan Matney, Joyce Sensmeier, Judith Warren, John Welton, and Bonnie Westra. 2014. A call to action: Engage in big data science. *Nursing Outlook* 62, 1 (2014), 64–65.

Classcentral. 2016. Data Science and Big Data—Free Online Courses. Retrieved from https://www.class-central.com/subject/data-science.

Kelly Clay. 2013. CES 2013: The Year of The Quantified Self? Retrieved from http://www.forbes.com/sites/kellyclay/2013/01/06/ces-2013-the-year-of-the-quantified-self/♯4cf4d2b55e74.

William S. Cleveland. 2001. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review* 69, 1 (2001), 21–26. DOI:http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x

CMIST. 2016. China Will Establish A Series of National Labs. Retrieved from http://news.sciencenet.cn/htmlnews/2016/4/344404.shtm.

CNSF. 2015. National Science Foundation China. Retrieved from http://www.nsfc.gov.cn/.

European Commission. 2014. Commission urges governments to embrace potential of big data. Retrieved from europa.eu/rapid/press-release_IP-14-769_en.htm.

Coursera. 2016. Coursera. Retrieved from www.coursera.org/data-science.

Kevin Crowston and Jian Qin. 2011. A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the Association for Information Science and Technology* 48, 10 (2011), 1–9.

CSC. 2012. Big data universe beginning to explode. Retrieved from http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode.

CSNSTC. 2009. Harnessing the Power of Digital Data for Science and Society: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Retrieved from https://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

DABS. 2016. Data Analytics Book Series. Retrieved from http://www.springer.com/series/15063.

DARPA. 2016. DARPA Xdata program. Retrieved from www.darpa.mil/program/xdata.

Data61. 2016. Data61. Retrieved from https://www.data61.csiro.au/.

DataRobot. 2016. DataRobot. Retrieved from https://www.datarobot.com/.

Datasciences.org. 2005. Homepage. Retrieved from www.datasciences.org.

Thomas H. Davenport and D. J. Patil. 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review* (2012), 70–76.

Jessica Davis. 2016. 10 Programming Languages And Tools Data Scientists Used. Retrieved from http://www.informationweek.com/devops/programming-languages/10-programming-languages-and-tools-data-scientists-use-now/d/d-id/1326034.

Devendra Desale. 2015. Top 30 Social Network Analysis and Visualization Tools. Retrieved from http://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html.

Vasant Dhar. 2013. Data science and prediction. *Communications of the ACM* 56, 12 (2013), 64–73.

Herman A. Dierick and Fabrizio Gabbiani. 2015. Drosophila neurobiology: No escape from 'Big Data' science. *Current Biology* 25, 14 (2015), 606–608.

Peter J. Diggle. 2015. Statistics: A data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178, 4 (2015), 793–813.

David Donoho. 2015. 50 years of Data Science. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

Bonnie J. Dorr, Craig S. Greenberg, Peter Fontana, Mark A. Przybocki, Marion Le Bras, Cathryn A. Ploehn, Oleg Aulov, Martial Michel, E. Jim Golden, and Wo Chang. 2015. The NIST data science initiative. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA'15)*. 1–10.

DSA. 2016. Data Science Association. Retrieved from http://www.datascienceassn.org/.

DSAA. 2014. IEEE/ACM/ASA International Conference on Data Science and Advanced Analytics. Retrieved from www.dsaa.co.

DSC. 2016a. College & University Data Science Degrees. Retrieved from http://datascience.community/colleges.

DSC. 2016b. The Data Science Community. Retrieved from http://datasciencebe.com/.

DSCentral. 2016. Data Science Central. Retrieved from http://www.datasciencecentral.com/.

DSE. 2015. Data Science and Engineering. Retrieved from http://link.springer.com/journal/41019.

DSJ. 2014. Data Science Journal.Retrieved from datascience.codata.org.

DSKD. 2007. Data Science and Knowledge Discovery Lab, UTS. Retrieved from http://www.uts.edu.au/research-and-teaching/our-research/quantum-computation-and-intelligent-systems/data-sciences-and.

David Ewing Duncan. 2009. *Experimental Man: What One Man's Body Reveals about His Future, Your Health, and Our Toxic World*. Wiley & Sons, New York.

Edx. 2016. EDX Courses. Retrieved from https://www.edx.org/course?search_query=data+science.

EMC. 2011. Data science revealed: A data-driven glimpse into the burgeoning new field. Retrieved from www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf.

EPJDS. 2012. EPJ Data Science. Retrieved from http://epjdatascience.springeropen.com/.

EU. 2014. EU Towards a Thriving Data-Driven Economy. Retrieved from https://ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy.

EU-DSA. 2016. The European Data Science Academy. Retrieved from edsa-project.eu.

EU-OD. 2016. The European Union Open Data Portal. Retrieved from https://open-data.europa.eu/.

Facebook. 2016. Facebook Data. Retrieved from https://www.facebook.com/careers/teams/data/.

James H. Faghmous and Vipin Kumar. 2014. A big data guide to understanding climate change: The case for theory-guided data science. *Big Data* 2, 3 (2014), 155–163.

Joshua Fairfielda and Hannah Shteina. 2014. Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics* 29, 1 (2014), 38–51.

Jack Faris, Evelyne Kolker, Alex Szalay, Leon Bradlow, Ewa Deelman, Wu Feng, Judy Qiu, Donna Russell, Elizabeth Stewart, and Eugene Kolker. 2011. Communication and data-intensive science in the beginning of the 21st century. *A Journal of Integrative Biology* 15, 4 (2011), 213–215.

Tom Fawcett. 2016. Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data* 3, 4 (2016), 249–266.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17, 3 (1996), 37–54.

William Finzer. 2013. The data science education dilemma. *Technology Innovations in Statistics Education* 7, 2 (2013).

Geoffrey Fox, Siddharth Maini, Howard Rosenbaum, and David J. Wild. 2015. Data science and online education. In *Proceedings of the 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom'15)*. 582–587.

Peter Fox and James Hendler. 2014. The science of data science. *Big Data* 2, 2 (2014), 68–70.

Molly Galetto. 2016. Top 50 Data Science Resources. Retrieved from http://www.ngdata.com/top-data-science-resources/?.

GEO. 2016. Gene Expression Omnibus. Retrieved from http://www.ncbi.nlm.nih.gov/geo/.

Deepak Ghodke. 2015. Bye Bye 2015: What lies ahead for BI. Retrieved from http://www.ciol.com/bye-bye-2015-what-lies-ahead-for-bi/.

Github. 2016a. Data Science Colleges. Retrieved from https://github.com/ryanswanstrom/awesome-datascience-colleges.

Github. 2016b. List of Recommender Systems. Retrieved from https://github.com/grahamjenson/list_of_recommender_systems.

Michael Gold, Ryan McClarren, and Conor Gaughan. 2013. The lessons Oscar taught us: Data science and media & entertainment. *Big Data* 1, 2 (2013), 105–109.

Google. 2016a. Google Bigquery and Cloud Platform. Retrieved from https://cloud.google.com/bigquery/.

Google. 2016b. Google Cloud Prediction API. Retrieved from https://cloud.google.com/prediction/docs/.

Google. 2016c. Google Online Open Education. Retrieved from https://www.google.com/edu/openonline/.

Google. 2016d. Google Trends. (2016). https://www.google.com.au/trends/explore#q=datalyticsz=Etc Retrieved on 14 Novermber 2016.

Google. 2016e. Open Mobile Data. Retrieved from https://console.developers.google.com/storage/browser/openmobiledata_public/.

Beijing Municipal Government. 2016. Beijing Big Data and Cloud Computing Development Action Plan. Retrieved from http://zhengwu.beijing.gov.cn/gh/dt/t1445533.htm.

China Government. 2015. China Big Data. Retrieved from http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

Matthew J. Graham. 2012. The art of data science. In *Astrostatistics and Data Mining,* Springer Series in Astrostatistics, Vol. 2. 47–59.

Jim Gray. 2007. eScience—A Transformed Scientific Method. Retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.

GTD. 2016. Global Terrorism Database. Retrieved from https://www.start.umd.edu/gtd/.

Akash Gupta, Ahmet Cecen, Sharad Goyal, Amarendra K. Singh, and Surya R. Kalidindi. 2015. Structure-property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system. *Acta Materialia* 91 (2015), 239–254.

David J. Hand. 2015. Statistics and computing: The genesis of data science. *Statistics and Computing* 25, 4 (2015), 705–711.

Hardin. 2016. Github. Retrieved from hardin47.github.io/DataSciStatsMaterials/.

Johanna Hardin, Roger Hoerl, Nicholas J. Horton, and Deborah Nolan. 2015. Data science in statistics curricula: Preparing students to “Think with Data”. *The American Statistician* 69, 4 (2015), 343–353.

Harlan Harris, Sean Murphy, and Marck Vaisman. 2013. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O’Reilly Media.

Benjamin T. Hazena, Christopher A. Booneb, Jeremy D. Ezellc, and L. Allison Jones-Farmer. 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics* 154 (2014), 72–80.

Tony Hey, Stewart Tansley, and Kristin Tolle (Eds.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

Tony Hey and Anne Trefethen. 2003. *The Data Deluge: An e-Science Perspective*. John Wiley & Sons, Ltd, 809–824.

HLSG. 2010. Final report of the high level expert group on scientific data. http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707.

HLSG. 2014. An RDA Europe Report. Retrieved from http://www.e-nformation.ro/wp-content/uploads/2014/12/TheDataHarvestReport_-Final.pdf.

Horizon. 2014. European Commission Horizon 2020 Big Data Private Public Partnership. Retrieved from http://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies.

Peter J. Huber. 2011. *Data Analysis: What Can Be Learned From the Past 50 Years*. John Wiley & Sons.

IASC. 1977. International Association for Statistical Computing. (1977). http://www.iasc-isi.org/.

IBM. 2010. Capitalizing on Complexity. Retrieved from http://www-935.ibm.com/services/us/ceo/ceostudy2010/multimedia.html.

IBM. 2016a. IBM Analytics and Big Data. Retrieved from http://www.ibm.com/analytics/us/en/orhttp://www-01.ibm.com/software/data/bigdata/.

IBM. 2016b. What is a Data Scientist? Retrieved from http://www-01.ibm.com/software/data/infosphere/data-scientist/.

IDA. 2014. International Institute of Data & Analytics. Retrieved from www.datasciences.org.

IEEEBD. 2014. IEEE Big Data Initiative. (2014). http://bigdata.ieee.org/.

IFSC-96. 1996. Data Science, Classification, and Related Methods. Retrieved from http://d-nb.info/955715512/04.

IJDS. 2016. International Journal of Data Science. (2016). http://www.inderscience.com/jhome.php?jcode=ijds.

IJRDS. 2017. International Journal of Research on Data Science. Retrieved from http://www.sciencepublishinggroup.com/journal/index?journalid=310.

INFORMS. 2014. Candidate Handbook. Retrieved from https://www.informs.org/Certification-Continuing-Ed/Analytics-Certificati on/Candidate-Handbook.

INFORMS. 2016. Institute for Operations Research and the Management Sciences. Retrieved from https://www.informs.org/.

Shuichi Iwata. 2008. Scientific “agenda” of data science. *Data Science Journal* 7, 5 (2008), 54–56.

H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. *Communications of the ACM* 57, 7 (2014), 86–94.

H. V. Jagadish. 2015. Big data and science: Myths and reality. *Big Data Research* 2, 2 (2015), 49–52.

JDS. 2002. Journal of Data Science. Retrieved from http://www.jds-online.com/.

JDSA. 2015. International Journal of Data Science and Analytics (JDSA). Retrieved from http://www.springer.com/41060.

JFDS. 2016. The Journal of Finance and Data Science. Retrieved from http://www.keaipublishing.com/en/journals/the-journal-of-finance-and-data-science/.

Kaggle. 2016. Kaggle Competition Data. Retrieved from https://www.kaggle.com/competitions.

Surya R. Kalidindi. 2015. Data science and cyberinfrastructure: Critical enablers for accelerated development of hierarchical materials. *International Materials Reviews* 60, 3 (2015), 150–168.

KDD89. 1989. IJCAI-89 Workshop on Knowledge Discovery in Databases. Retrieved from http://www.kdnuggets.com/meetings/kdd89/index.html.

KDnuggets. 2015. Visualization Software. Retrieved from http://www.kdnuggets.com/software/visualization.html.

Kdnuggets. 2016. Kdnuggets. Retrieved from http://www.kdnuggets.com/.

K Kelly. 2012. The quantified century. In *Quantified Self Conference*. Retrieved from http://quantifiedself.com/conference/Palo-Alto-2012.

Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, and et al. 2014. Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal* 2014 (2014), 18.

John King and Roger Magoulas. 2015. 2015 Data Science Salary Survey. Retrieved from http://duu86o6n09pv.cloudfront.net/reports/2015-data-science-salary-survey.pdf.

Ron Kohavi, Neal J. Rothleder, and Evangelos Simoudis. 2002. Emerging trends in business analytics. *Communications of the ACM* 45, 8 (2002), 45–48.

AMP Lab. 2016. MLBase. Retrieved from http://mlbase.org/.

Alexandros Labrinidis and H. V. Jagadish. 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2032–2033.

Douglas Laney. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical Report, META Group.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google flu: Traps in big data analysis. *Science* 343 (2014), 1203–1205.

LDC. 2016. Linguistic Data Consortium. Retrieved from https://www.ldc.upenn.edu/about.

LinkedIn. 2016. LinkedIn Jobs. Retrieved from https://www.linkedin.com/jobs/data-scientist-jobs.

Mike Loukides. 2011. *The Evolution of Data Products*. O'Reilly, Cambridge.

Mike Loukides. 2012. *What is Data Science?* O'Reilly Media, Sebastopol, CA. http://radar.oreilly.com/2010/06/what-is-data-science.htmldata-scientists.

Andrea Manieri, Steve Brewer, Ruben Riestra, Yuri Demchenko, Matthias Hemmje, Tomasz Wiktorski, Tiziana Ferrari, and Jrmy Frey. 2015. Data science professional uncovered: How the EDISON project will contribute to a widely accepted profile for data scientists. In *Proceedings of the 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom'15)*. 588–593.

Kate Matsudaira. 2015. The science of managing data science. *Communications of the ACM* 58, 6 (2015), 44–47.

McKinsey. 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

Claire Cain Miller. 2013. Data science: The numbers of our lives. *New York Times* Retrieved from http://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html?pagewanted=all&_r=0.

Arthur John Havart Morrell (Ed.). 1968. Information processing. In *Proceedings of IFIP Congress 1968*. Edinburgh, UK.

Peter Murray-Rust. 2007. Data-driven science: A scientist's view. In *NSF/JISC 2007 Digital Repositories Workshop*. http://www.sis.pitt.edu/repwkshop/papers/murray.pdf.

Peter Naur. 1968. 'Datalogy', the science of data and data processes. In *Proceedings of IFIP Congress 1968*, 1383–1387.

Peter Naur. 1974. *Concise Survey of Computer Methods*. Studentlitteratur, Lund, Sweden.

NCSU. 2007a. Institute for Advanced Analytics, North Carolina State University. Retrieved from http://analytics.ncsu.edu/.

NCSU. 2007b. Master of Science in Analytics, Institute for Advanced Analytics, North Carolina State University. Retrieved from http://analytics.ncsu.edu/.

Michael L. Nelson. 2009. Data-driven science: A new paradigm? *EDUCAUSE Review* 44, 4 (2009), 6–7.

NICTA. 2016. National ICT Australia. Retrieved from https://www.nicta.com.au/.

NIST. 2015. NIST Text Retrieval Conference Data. Retrieved from http://trec.nist.gov/data.html.

NSB. 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/.

NSF. 2007. US NSF07-28. Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

OECD. 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. Retrieved from https://www.oecd.org/sti/sci-tech/38500813.pdf.

OPENedX. 2016. OPENedX Online Education Platform. Retrieved from https://open.edx.org/.

Tim O'Reilly. 2005. What is Web 2.0. Retrieved from http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=3.

D. J. Patil. 2011. *Building Data Science Teams*. O'Reilly Media.

Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. 1993. Capability maturity model version 1.1. *IEEE Software* 10, 4 (1993), 18–27.

Gil Press. 2013. A Very Short History of Data Science. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/61ae3ebb69fd.

Xuesen Qian. 1991. Revisiting issues on open complex giant systems. *International Journal of Pattern Recognition and Artificial Intelligence* 4, 1 (1991), 5–8.

Xuesen Qian, Jingyuan Yu, and Ruwei Dai. 1993. A new discipline of science—The study of open complex giant system and its methodology. *Chinese Journal of Systems Engineering & Electronics.* 4, 2 (1993), 2–12.

RapidMiner. 2016. RapidMiner. (2016). https://rapidminer.com/.

Samantha Renae. 2011. Data analytics: Crunching the future. *Bloomberg Businessweek* (2011). September 8.

Solutions Review. 2016. Data Integration and Application Integration Solutions Directory. Retrieved from http://solutionsreview.com/data-integration/data-integration-solutions-directory/.

C. Rudin, D. Dunson, R. Irizarry, H. Ji, E. Laber, J. Leek, T. McCormick, Sherri Rose, C. Schafer, M. van der Laan, L. Wasserman, and L. Xue. 2014. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society. Retrieved from http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf American Statistical Association.

SAS. 2013. Big Data Analytics: An Assessment of Demand for Labour and Skills, 2012-2017. Retrieved from https://www.thetechpartnership.com/globalassets/pdfs/research-2014/bigdata_report_nov14.pdf Report. SAS/The Tech Partnership.

SAS. 2016. SAS Retrieved from http://www.sas.com/en_us/insights.html.

Tobias Schoenherr and Cheri Speier-Pero. 2015. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics* 36, 1 (2015), 120–132.

SIAM. 2016. SIAM career center. (2016). http://jobs.siam.org/home/.

Christoph Siart, Simon Kopp, and Jochen Apel. 2015. The interface between data science, research assessment and science support—Highlights from the German perspective and examples from Heidelberg University. In *Proceedings of the 2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI'15)*. 472–476.

Silk. 2016. Data Science University Programs. Retrieved from http://data-science-university-programs.silk.co/.

Larry Smarr. 2012. Quantifying your body: A how-to guide from a systems biology perspective. *Biotechnology Journal* 7, 8 (2012), 980–991. DOI:http://dx.doi.org/10.1002/biot.201100495

F. Jack Smith. 2006. Data science as an academic discipline. *Data Science Journal* 5 (2006), 163–164.

SSDS. 2015. Springer Series in the Data Sciences. Retrieved from http://www.springer.com/series/13852.

Stanford. 2014. Stanford Data Science Initiatives, Stanford University. Retrieved from https://sdsi.stanford.edu/.

Thomas R. Stewart and Claude McMillan, Jr. 1987. Descriptive and prescriptive models for judgment and decision making: Implications for knowledge engineering. In *Expert Judgment and Expert Systems*, Jeryl L. Mumpower, Ortwin Renn, Lawrence D. Phillips, and V. R. R. Uppuluri (Eds.). Springer-Verlag, London, 305–320.

Michael Stonebraker, Sam Madden, and Pradeep Dubey. 2013. Intel 'big data' science and technology center vision and execution plan. *SIGMOD Record* 42, 1 (2013), 44–49.

Alma Swan and Sheridan Brown. 2008. The skills, role career structure of data scientists curators: Assessment of current practice future needs. (2008). Technical Report. University of Southampton.

Melanie Swan. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data* 1, 2 (2013), 85–99.

Technavio. 2016. Top 10 Healthcare Data Analytics Companies. Retrieved from http://www.technavio.com/blog/top-10-healthcare-data-analytics-companies.

TFDSAA. 2013. IEEE Task Force on Data Science and Advanced Analytics. Retrieved from http://dsaatf.dsaa.co/.

TOBD. 2015. IEEE Transactions on Big Data. Retrieved from https://www.computer.org/web/tbd.

Predictive Analytics Today. 2016. 29 Data Preparation Tools and Platforms. Retrieved from http://www.predictiveanalyticstoday.com/data-preparation-tools-and-platforms/.

John W. Tukey. 1962. The future of data analysis. *The Annals of Mathematical Statistics* 33, 1 (1962), 1–67.

John W. Tukey. 1977. *Exploratory Data Analysis*. Pearson.

Tutiempo. 2016. Global Climate Data. Retrieved from http://en.tutiempo.net/climate.

UCI. 2016. UCI Machine Learning Repository. Retrieved from archive.ics.uci.edu/ml/.

Udacity. 2016. Udacity Courses. Retrieved from https://www.udacity.com/courses/data-science.

Udemy. 2016. Udemy Courses. Retrieved from https://www.udemy.com/courses/search/?ref=home&src=ukw&q=data+science&lang=en.

UK. 2016. UK Big Data. Retrieved from http://www.rcuk.ac.uk/research/infrastructure/big-data/.

UK-HM. 2012. UK HM Government. Retrieved from http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf.

UK-OD. 2016. UK Open Data. Retrieved from http://data.gov.uk/.

UMichi. 2015. Michigan Institute For Data Science, University of Michigan. Retrieved from http://midas.umich.edu/.

UN. 2010. United Nation Global Pulse Projects. Retrieved from http://www.unglobalpulse.org/.

US-OD. 2016. US Government Open Data. Retrieved from https://www.data.gov/.

USD2D. 2016. US National Consortium for Data Science. Retrieved from data2discovery.org.

USDSC. 2016. US Degree Programs in Analytics and Data Science. Retrieved from http://analytics.ncsu.edu/?page_id=4184.

USNSF. 2012. US Big Data Research Initiative. Retrieved from http://www.nsf.gov/cise/news/bigdata.jsp.

UTS. 2011. Master of Analytics (Research) and Doctor of Philosophy Thesis: Analytics, Advanced Analytics Institute, University of Technology Sydney. Retrieved from http://www.uts.edu.au/research-and-teaching/our-research/advanced-analytics-institute/education-and-research-opportuniti-1.

UTSAAI. 2011. Advanced Analytics Institute, University of Technology Sydney. Retrieved from https://analytics.uts.edu.au/.

David van Dyk, Montse Fuentes, Michael I. Jordan, Michael Newton, Bonnie K. Ray, Duncan Temple Lang, and Hadley Wickham. 2015. ASA Statement on the Role of Statistics in Data Science. Retrieved from http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/.

Vast. 2016. Visual Analytics Community. Retrieved from http://vacommunity.org/HomePage.

Dan Vesset, Benjamin Woo, Henry D. Morris, Richard L. Villars, Gard Little, Jean S. Bozman, Lucinda Borovick, Carl W. Olofson, Susan Feldman, Steve Conway, Matthew Eastwood, and Natalya Yezhkova. 2012. Worldwide Big Data Technology and Services 2012-2015 Forecast. IDC.

Ana Viseu and Lucy Suchman. 2010. *Wearable Augmentations: Imaginaries of the Informed Body*. Berghahn Books, New York, 161–184.

Whitehouse. 2015. The White House Names Dr. D. J. Patil as the First U.S. Chief Data Scientist. Retrieved from https://www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist.

Wikipedia. 2016a. Comparison of Cluster Software. Retrieved from https://en.wikipedia.org/wiki/Comparison_of_cluster_software.

Wikipedia. 2016b. Informatics. (2016). https://en.wikipedia.org/wiki/Informatics.

Wikipedia. 2016c. List of Reporting Software. Retrieved from https://en.wikipedia.org/wiki/List_of_reporting_software.

WIRED. 2014. How Europe can Seize the Starring Role in Big Data. Retrieved from www.wired.com/insights/2014/09/europe-big-data/.

Gary Wolf. 2012. The data-driven life. *New York Times*. Retrieved from www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html.

Jeff Wu. 1997. Statistics = Data Science? Retrieved from http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.

Yahoo. 2016. Yahoo Finance. Retrieved from finance.yahoo.com.

Nathan Yau. 2009. Rise of the Data Scientist. Retrieved from http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/.

Chris Yiu. 2012. The Big Data Opportunity. Retrieved from http://www.policyexchange.org.uk/images/publications/thepportunity.pdf.

Bin Yu. 2014. IMS presidential address: Let us own data science. *IMS Bulletin Online* (2014). Oct. 1, 2014.

Received June 2016; revised November 2016; accepted April 2017