

Data Science: Nature and Pitfalls

Longbing Cao

Advanced Analytics Institute, University of Technology Sydney, Australia

Abstract—Data science is creating very exciting trends as well as significant controversy. A critical matter for the healthy development of data science in its early stages is to deeply understand the nature of data and data science, and to discuss the various pitfalls. These important issues motivate the discussions in this article.

Keywords—Data science, big data, analytics, advanced analytics, big data analytics

I. INTRODUCTION

The era of analytics [57], data science [58], and big data [44] has driven substantial governmental, industrial, and disciplinary interest, goal and strategy transformation as well as a paradigm shift in research and innovation. This has resulted in significant new opportunities and prospects becoming available which were not previously possible, while an increasing and overwhelming amount of fanfare and hype has spread across multiple domains, areas and events.

In reviewing the related initiatives, progress, and status of data science, analytics and big data [9], and the diversified discussions about prospects, challenges and directions [11], the controversy caused by the potential conflict of these various elements becomes clear. This review discloses the need for deep discussions about the nature and pitfalls of data science, the clarification of fundamental concepts and myths, and the demonstration of the intrinsic characteristics and opportunities of data science.

This paper thus focuses on discussing the two fundamental issues in data science: the nature and the pitfalls of data science. Addressing the first highlights the status, intrinsic factors, characteristics, and features of the era of data science and analytics, as well as the challenges and opportunities for new generation innovation, research and disciplinary development. The second summarizes common pitfalls about the concepts of data science, data volume, infrastructure, analytics, and capabilities and roles. Building on the above discussions, the concepts and possible future directions of data science are then presented.

II. FEATURES OF THE DATA SCIENCE ERA

Drawing a picture of the features and characteristics of the era of data science is critical and challenging. We explore this from the perspective of the transformation and paradigm shift caused by data science, and discuss the core driving forces, as well as the status of a number of typical issues confronting the data science field.

A. The Era's Transformation and Paradigm Shift

The emergence of the era of data science and analytics can be highlighted by three key indicators: a significant *disciplinary paradigm shift*, *technological transformation*, and *innovative production*.

- **Disciplinary paradigm shifting:** The shifting of data-centric disciplinary paradigms from one to another.
- **Technological transformation:** The upgrading of data technology from one generation to another.
- **Innovative production:** The innovation of technical and practical data products.

These are discussed below.

The disciplinary paradigm shift of data-oriented and data-centric research, innovation and profession can be embodied by such aspects as:

- From data analysis to data analytics;
- From descriptive analytics to deep analytics;
- From data analytics to data science;

The disciplinary paradigm shift promotes data-related technological transformation by such means as follows:

- From large-scale data to big data;
- From business operational systems to business analytical systems;
- From World Wide Web to Wisdom web;
- From Internet to Internet of Everything (incl. Internet of Things, mobile network, social network);

Innovative production in the data and analytics areas can be represented by typical indicators such as the following:

- From digital economy to data economy;
- From closed government to open government;
- From e-commerce to online business;
- From telephone to smart phone;
- From Internet to mobile and social network.

B. Data-centric Driving Forces

The transformation and paradigm shift of data-oriented discipline, technologies and production are driven by core forces including data-enabled opportunities, data-related ubiquitous factors, and various complexities and intelligences embedded in data-oriented production and products.

Ubiquitous data-oriented factors include data, behavior, complexity, intelligence, service, and opportunities.

- **Data is ubiquitous:** Involving historical, real-time and future data;
- **Behavior is ubiquitous:** Bridging the gaps between the physical world and the data world;

- Complexity is ubiquitous: The types and extent of complexity differentiate one data system from another;
- Intelligence is ubiquitous: Diversified intelligences are embedded in a data system;
- Service is ubiquitous: Data services present in various forms and domains;
- Opportunities are ubiquitous: Data enables enormous opportunities.

Data-enabled opportunities, also called *X-opportunities*, are overwhelming and extend from research, innovation, education, and government to economy. We briefly elaborate on them below.

- Research opportunities: Inventing data-focused breakthrough theories and technologies;
- Innovation opportunities: Developing data-based cutting-edge services, systems and tools;
- Education opportunities: Innovating data-oriented courses and training;
- Government opportunities: Enabling data-driven government decision-making and objectives;
- Economic opportunities: Fostering data economy, services, and industrialization;
- Lifestyle opportunities: Promoting data-enabled smarter living and smarter cities;
- Entertainment opportunities: Creating data-driven entertainment activities, networks, and societies.

A data science problem is a complex system [46], [8] in which comprehensive system complexities, also called *X-complexities* [11], are embedded. These comprise complexities of data characteristics, behavior, domain, society (social complexity), environment, learning, and decision-making.

- Data complexity: Embodied by such factors as comprehensive data circumstances and characteristics;
- Behavior complexity: Demonstrated by such aspects as individual and group activities, evolution, utility, impact and change;
- Domain complexity: Represented by such aspects as domain factors, processes, norms, policies, knowledge and domain expert engagement in problem-solving;
- Social complexity: Indicated by such aspects as social networking, community formation and divergence, sentiment, the dissemination of opinion and influence, and other social issues such as trust and security;
- Environment complexity: Capturing such aspects as contextual factors, interactions with systems, changes, and uncertainty.
- Learning complexity: Including the development of appropriate methodologies, frameworks, processes, models and algorithms, and theoretical foundation and explanation;
- Decision complexity: Involving such issues as the methods and forms of deliverables, communications and decision-making actions.

In a complex data science problem, ubiquitous intelligence, also called *X-intelligence* [11], is often demonstrated and has

to be incorporated and synergized [8] in problem-solving processes and systems.

- Data intelligence: Highlighting the interesting information, insights and stories hidden in data about business problems and driving forces.
- Behavior intelligence: Demonstrating the insights of activities, processes, dynamics, impact and trust of individual and group behaviors by human and action-oriented organisms.
- Domain intelligence: Domain values and insights emerge from involving domain factors, knowledge, meta-knowledge, and other domain-specific resources.
- Human intelligence: Contributions made by the empirical knowledge, beliefs, intentions, expectations, critical thinking and imaginary thinking of human individual and group actors.
- Network intelligence: Intelligence created by the involvement of networks, web, and networking mechanisms in problem comprehension and problem-solving.
- Organizational intelligence: Insights and contributions created by the involvement of organization-oriented factors, resources, competency and capabilities, maturity, evaluation and dynamics.
- Social intelligence: Contributions and values generated by the inclusion of social, cultural, and economic factors, norms and regulation.
- Environmental intelligence: This may be embodied through other intelligences specific to the underlying domain, organization, society, and actors.

The above data-oriented and data-driven factors, complexities, intelligences and opportunities constitute the nature and characteristics of data science, and drive the evolution and dynamics of data science problems.

C. Data DNA

In the biological domain, DNA is a molecule that carries genetic instructions that are uniquely valuable to the biological development, functioning and reproduction of humans and all known living organisms. As a result of data quantification, data is everywhere, and is present in the public Internet, Internet of Things, sensor networks, socio-cultural, economic and geographical repositories and quantified personalized sensors, including mobile, social, living, entertaining and emotional sources. This forms the “datalogical” constituent: “data DNA”, which plays a critical role in data organisms and performs a similar function to biological DNA in living organisms.

Definition 1 (Data DNA). Data DNA is the datalogical “molecule” of data, consisting of fundamental and generic constituents: entity (*E*), property (*P*) and relationship (*R*). Here “datalogical” means that data DNA plays a similar role in data organisms as biological DNA plays in living organisms. Entity can be an object, an instance, a human, an organization, a system, or a part of a sub-system of a system. Property refers to the attributes that describe an entity.

Relationship corresponds to (1) entity interactions, and (2) property interactions, including property value interactions.

Entity, property and relationship present different characteristics in terms of quantity, type, hierarchy, structure, distribution and organization. A data-intensive application or system is often composed of a large number of diverse entities, each of which has specific properties, and different relationships are embedded within and between properties and entities. From the very lowest level to the very highest level, data DNA presents heterogeneity and hierarchical couplings across levels. On each level, it maintains *consistency* (inheritance of properties and relationships) as well as *variations* (mutations) across entities, properties and relationships, while *personalized characteristics* are supported for each individual entity, property and relationship.

For a given data, its entities, properties and relationships are instantiated into diverse and domain-specific forms, which carry most of the data ecological and genetic information in data generation, development, functioning, reproduction, and evolution. In the data world, *data DNA* is embedded in the whole body of personal [65] and non-personal data organisms, and in the generation, development, functioning, management, analysis and use of all data-based applications and systems.

Data DNA drives the evolution of a data-intensive organism. For example, university data DNA connects the data of students, lecturers, administrative systems, corporate services and operations. The student data DNA further consists of academic, pathway, library access, online access, social media, mobile service, GPS, and Wifi usage data. Such student data DNA is both steady and evolving.

In complex data, data DNA is embedded within various X-complexities (see detailed discussions in [11] and [8]) and ubiquitous X-intelligence (more details in [11] and [8]) in a data organism. This makes data rich in content, characteristics, semantics and value, but challenging in acquisition, preparation, presentation, analysis and interpretation.

D. Data Quality

Data science tasks involve roles and follow processes different from more generalized IT projects, since data science and analytics works tend to be creative, intelligent, exploratory, non-standard, unautomated and personalized, and have the objective of discovering evidence and indicators for decision-making actions. They inevitably involve quality issues such as data validity, veracity, variability and reliability, and social issues such as privacy, security, accountability and trust, which need to be taken into account in data science and analytics.

Data quality is a critical problem in data science and engineering. Given a data science problem, we should not assume that

- The data available or given is perfect,
- The data always generates good outcomes,
- The outputs (findings) generated are always good and meaningful, and
- The outcomes can always inform better decisions.

These assumption myths involve the quality of the data (input), the model, and the outcomes (output), in particular, validity, veracity, variability and reliability. We briefly discuss these aspects below.

Data and analytics *validity* determines whether a data model, concept, conclusion or measurement is well-founded and corresponds accurately to the data characteristics and real-world facts, making it capable of giving the right answer.

Similarly, data and analytics *veracity* determines the correctness and accuracy of data and analytics outcomes. Both validity and veracity also need to be checked from the perspectives of data content, representation, design, modeling, experiments, and evaluation.

Data and analytics *variability* is determined by the changing and uncertain nature of data, reflecting business dynamics (including the problem context and problem-solving purposes), and thus requires the corresponding analytics to adapt to the dynamic nature of the data. Due to the changing nature of data, the need to check the validity, veracity and reliability of the data used and analytics undertaken is thus highly important.

Data and analytics *reliability* refers to the consistency, redundancy, repeatability and trust properties of the data used, the analytic models generated, and the outcomes delivered on the data. Reliable data and analytics are not necessarily static. Making data analytics adaptive to the evolving, streaming and dynamic nature of data, business and decision requests is a critical challenge in data science and analytics.

E. Social Issues

Domain-specific data and business are embedded in social contexts and incorporated with social issues. Data science tasks typically involve such social issues as *privacy*, *security*, *accountability* and *trust* on data, modeling and deliverables, which we discuss below.

Data and analytics *privacy* addresses the challenge of collecting, analyzing, disseminating and sharing data and analytics while protecting personally identifiable or other sensitive information and analytics from improper disclosure. Protection technology, regulation and policies are required to balance protection and appropriate disclosure in the process of data manipulation.

Data and analytics *security* protects target objects from destructive forces and from the unwanted actions of unauthorized users, including improper use or disclosure, and not only addresses privacy issues but also other aspects beyond privacy, such as software and hardware backup and recovery. Data and analytics security also involves the development of regulating, political or legal mechanisms and systems to address such issues.

Data and analytics *accountability* refers to an obligation to comply with data privacy and security legislation, and to report, explain, trace and identify the data manipulated and analytics conducted to maintain the transparency and traceability, liability and warranty of both measurement and results, as well as the efficacy and verifiability of analytics and protection.

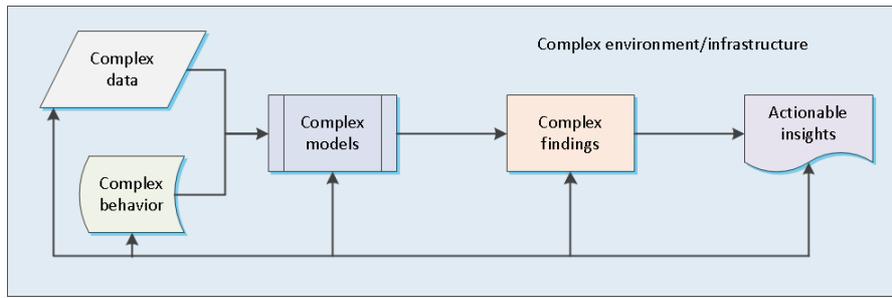


Fig. 1. The extreme data challenge.

Data and analytics *trust* refers to the belief in the reliability, truth or ability of data and analytics to achieve the relevant goals. This involves the development of appropriate technology, social norms, ethical rules, or legislation to ensure, measure and protect trust in the data and analytics used and confidence in the corresponding outcomes and evaluation of analytics.

F. The Extreme Challenge

Different types and levels of analytical problems trouble the existing knowledge base, and we are especially challenged by the complex problems in complex data and in complex environments. Our particular focus on data science research and innovation concerns the type of scenario we call an *extreme data challenge* in data science and analytics. The *extreme data challenge* as illustrated in Fig. 1, seeks to discover and deliver complex knowledge in complex data taking into account complex behavior within a complex environment to achieve actionable insights that will inform and enable decision action-taking in complex business problems that cannot be better handled by other methods.

The critical future directions of data science research and innovation in this case are focused on

- *Complex data* with complex characteristics and various complexities (as discussed above on data complexity and intelligence; for more, see [11] and [8]),
- *Complex behaviors* with complex relationships and dynamics (as discussed above on behavior complexity and intelligence; for more, see [11] and [8]),
- *Complex environment* in which complex data and behaviors are embedded and interacted with (as discussed above on domain-specific, organizational, social and environmental complexities and intelligence; for more, see [11] and [8]),
- *Complex models* to address the data and behavior complexities in complex environment (as discussed about learning complexities and decision complexities),
- *Complex findings* to uncover hidden but technically interesting and business-friendly observations, indicators or evidence, statements or presentations, and
- *Actionable insights* to evidence the next best or worst situation and inform the optimal strategies that should be taken to support effective business decision-making (see

more discussion on actionability in [6], [13].

Many real-life problems fall into this level of complexities and challenges, as shown in the extreme data challenge, and they have not been addressed well; for example,

- Understanding group behaviors by multiple actors where there are complex interactions and relationships, such as in the manipulation of large-scale cross-capital markets pool by internationally collaborative investors [12], each of whom plays a specific role by connecting information from the underlying markets, social media, other financial markets, socio-economic data and policies [7];
- Predicting local climate change and effect by connecting local, regional and global climate data, geographical data, agricultural data and other information [28].

III. DISCIPLINARY DEVELOPMENT OF DATA SCIENCE

In this section, we present a status summary of the disciplinary development of data science by reviewing the development gaps between the potential that data may have and the state-of-the-art capabilities to fulfill such potential, the research map of data science, and the course framework of data science.

A. Data-to-Capability Development Gaps

The rapid increase in big data unfortunately does not only present opportunities. As discussed in [6] and [13], there are significant gaps between what we are of and what we are capable of understanding. An empirical observation of the *data development gaps* between (1) the growth of *data potentials* and (2) the *state-of-the-art capabilities* is shown in Fig. 2. Such gaps have increased in the last 10 years and especially recently, due to the imbalance between potential exponential increase and progressive state-of-the-art capability development. We illustrate several such gaps below.

- Gap between data availability and currently understandable data level, scale and degree;
- Gap between data complexities and currently available analytics theories and tools;
- Gap between data complexities and currently available technical capabilities;
- Gap between possible values and impact and currently achievable outcomes and benefits;

- Gap between organizational needs and currently available talents/data scientists;
- Gap between potential opportunities and current outcomes and benefits achievable.

Such growth gaps are driven by critical challenges for which there is a shortage of effective theories and tools. For example, a typical challenge in complex data concerns intrinsic complex coupling relationships and heterogeneity, forming non-IID data [7], which cannot be simplified in such a way that they can be handled by classic IID learning theories and systems. Other examples include the real-time learning of large-scale online data, such as learning shopping manipulation and making real-time recommendations on extremely high frequency data in the “11-11” shopping seasons launched by Alibaba, or identifying suspects in an extremely imbalanced and multi-source data and environment such as fraud detection in high frequency marketing trading. Other challenges are high invisibility, high frequency, high uncertainty, high dimensionality, dynamic nature, mixed sources, online learning at the web scale, and developing human-like thinking.

B. Research Map of Data Science

The way to explore the fundamental challenges and innovative opportunities facing big data and data science is to conduct problem-, data-, and goal-driven discovery.

- *Problem-driven discovery*: This requires understanding the intrinsic nature, characteristics, complexities, and boundaries of the problem, and then analyzing the gaps between the problem complexities and the existing capability set. This gap analysis is critical for original research and breakthrough scientific discovery.
- *Goal-driven discovery*: This requires understanding the business, technical and decision goals to be achieved by understanding the problem, and conducting gap analysis of what has been implemented and achieved and what is expected to be achieved.
- *Data-driven discovery*: This requires understanding the data characteristics, complexities and challenges in data, and the gaps between the nature of a problem and the data capabilities. Due to the limitations of existing data systems, projection from the underlying physical world where the problem sits to the data world where the problem is datafied may be biased, dishonest, or highly manipulated. As a result, the data does not completely capture the problem and thus cannot create a full picture of the through any type of data exploration.

There are two ways to explore major research challenges: one is to summarize what concerns the relevant communities, and the other is to scrutinize the potential issues arising from the intrinsic complexities and nature of data science problems as complex systems [8], [8]. Taking the first approach, we can obtain a picture of the main research challenges by summarizing the main topics and issues in the statistics communities [16], [67], [63], informatics and computing communities [52], [11], vendors [56], government initiatives [61], [60], [19], [20],

[59] and research institutions [62], [47], which focus on data science and analytics. The second approach is much more challenging, as it requires us to explore the unknown space of the complexities and comprehensive intelligence in complex data problems.

Below, we list some of the main challenges confronting the data science community in addressing the big data complexities represented by key topics in the data A-Z (see Section IV-A). We categorize the challenges facing domain-specific data applications and problems in terms of the following major areas:

- Challenges in data/business understanding,
- Challenges in mathematical and statistical foundations,
- Challenges in X-analytics and data/knowledge engineering,
- Challenges in data quality and social issues,
- Challenges in data value, impact and usability,
- Challenges in data-to-decision and actions.

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

- Behavior and event processing,
- Data storage and management systems,
- Data quality enhancement,
- Data modeling, learning and mining,
- Deep analytics, learning and discovery,
- Simulation and experimental design,
- High-performance processing and analytics,
- Analytics and computing architectures and infrastructure,
- Networking, communication and interoperation.

C. Course Framework of Data Science

The goal of data science and analytics education is to train and generate the data and analytics knowledge and proficiency required to manage the capability and capacity gaps in the creation of a data science profession [64], [42], and to achieve the goals of data science innovation and the data economy. Accordingly, different levels of education and training are necessary, from attending public courses, corporate training, and undergraduate courses, to joining a master of data science and/or PhD in data science program.

Public courses are designed for the general community, to lift their understanding, skills, profession and specialism in data science through multi-level short courses. They range from basic courses to intermediary and advanced courses. The knowledge map consists of such components as data science, data mining, machine learning, statistics, data management, computing, programming, system analysis and design, and modules related to case studies, hands-on practices, project management, communication, and decision support.

Corporate training and workshops are customized to upgrade and foster corporate thinking, knowledge, capability and practices for entire enterprise innovation and raising productivity. This involves offering courses and workshops for the workforce, from senior corporate executives to business

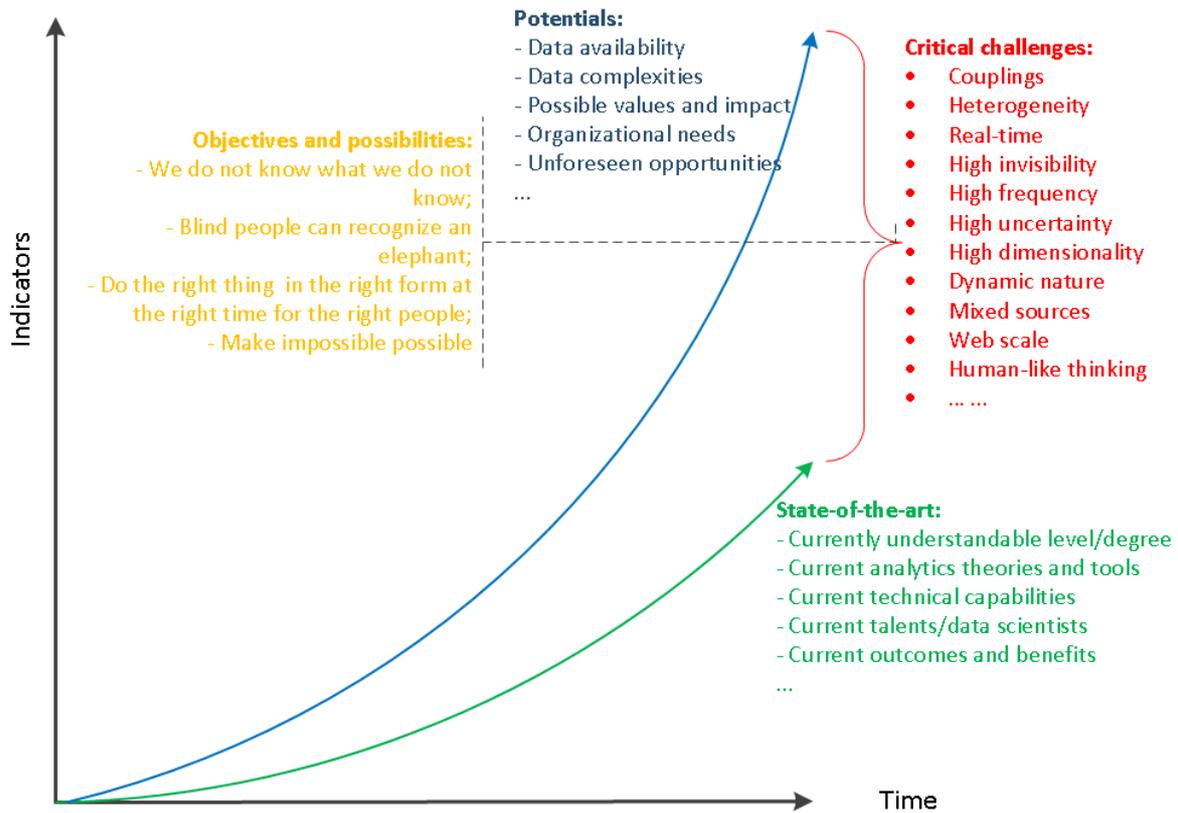


Fig. 2. Critical development gaps between data potentials and state-of-the-art capabilities.

owners, business analysts, data modelers, data scientists, data engineers, and deployment and enterprise strategists. Such courses cover scope and topics such as data science, data engineering, analytics science, decision science, data and analytics software engineering, project management, communications, and case management.

Undergraduate courses may be offered on either a general data science basis that focuses on building data science foundations and data and analytics computing, or specific areas such as data engineering, predictive modeling, and visualization. Double degrees or majors may be offered to train professionals who will gain knowledge and abilities across disciplines such as business and analytics, or statistics and computing.

Master of data science and analytics aims to train specialists and foster the talent of those who have the capacity to conduct a deep understanding of data and undertake analytics tasks in data mining, knowledge discovery and machine learning-based advanced analytics. Interdisciplinary experts may be trained from those who have a solid foundation in statistics, business, social science or other specific disciplines and are able to integrate data-driven exploration technologies with disciplinary expertise and techniques. A critical area in which data science and analytics should be incorporated is the classic master of business administration course. This is where new generation business leaders can be trained for the new economy and a global view of economic growth.

PhD in data science and analytics aims to train high level talent and specialists who have independent thinking, leadership, research, innovation and better practices for theoretical innovation to manage the significant knowledge and capability gaps, and for substantial economic innovation and raising productivity. Interdisciplinary research is encouraged to train leaders who have a systematic and strategic understanding of the what, how and why of data and economic innovation.

Fig. 3 shows the level, objective, capability set and outcomes of hierarchical data science and analytics education and training.

IV. DATA SCIENCE AS A NEW SCIENCE

So, what makes data science a new science? In this section, we discuss *data A-Z*, which may be used to capture every aspect of data science to form a data science ontological system, *the concept of data science*, which is built on the above discussions about the features and disciplinary development of data science, and *the future of data science*.

A. Data A-Z

In the big data community, multiple Vs are typically used to describe what constitutes big data, i.e., the characteristics, challenges and opportunities of big data. They include the volume (size), velocity (speed), variety (diversity), veracity (quality and trust), value (insight), visualization, and variability (formality) of data.

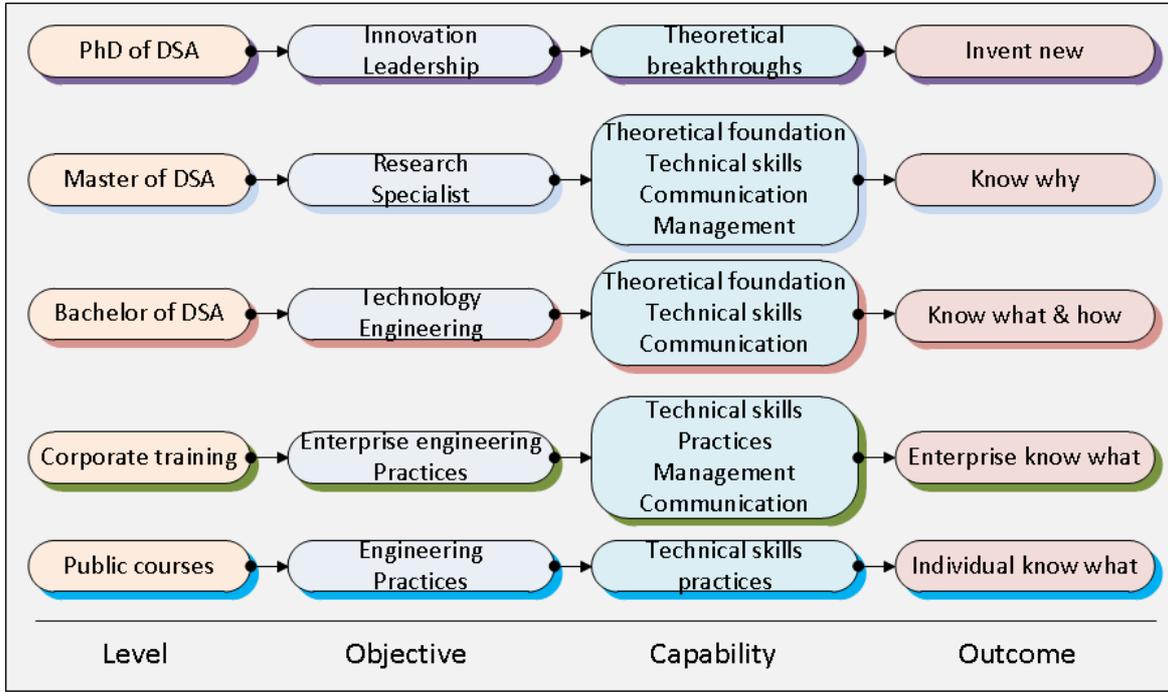


Fig. 3. Data science course framework.

In fact, these big Vs cannot describe a complete picture of big data, and cannot capture the field of data science. Therefore, it is very valuable to build a *data A-Z dictionary* to represent and capture the intrinsic comprehensive but diverse aspects, characteristics, challenges, domains, tasks, processes, purposes, applications and outcomes of, or related to, data. To this end, we list a sample sequence of data science keywords:

Actionability/Adaptation, Behavior/Boosting, Causality/Change, Dimensionality/Divergence, Embedding/Ethics, Fusion/Forecasting, Governance/Generalization, Heterogeneity/Hashing, Integrity/Inference, Join/Jungle, Kernelization/Knowledge, Linkage/Learning, Metrology/Migration, Normalization/Novelty, Optimization/Outlier, Privacy/Provenance, Quality/Quantity, Relation/Regularization, Scalability/Sparsity, Transformation/Transfer, Utility/Uncertainty, Variation/Visibility, Wrangling/Weighting, X – analytics/X – informatics, Yield/Yipit, Zettabyte/Zenit. (1)

It is notable that such a data A-Z ontology probably covers most of the topics of interest to major data science communities. The exercise of constructing data A-Z can substantially deepen and broaden the understanding of intrinsic data characteristics, complexities, challenges, prospects and opportunities [30].

B. What Is Data Science

Generally speaking, *data science is the science of data*, which concerns the study of data. There are different ways to define what data science is: it may be object-focused, process-based, and/or discipline-oriented [11].

Definition 2 (Data Science). *From the process perspective, data science is a systematic approach to ‘think with wisdom’, ‘understand domain’, ‘manage data’, ‘compute with data’, ‘mine on knowledge’, ‘communicate with stakeholders’, ‘deliver products’, and ‘act on insights’.*

A process-based data science formula is accordingly given below:

$$\text{data science} = \text{think} + \text{understand} + \text{manage} + \text{compute} + \text{mine} + \text{communicate} + \text{deliver} + \text{act} \quad (2)$$

In contrast, *data analytics understands data and its underlying business, discovers knowledge, delivers actionable insights, and enable decision-making*. From this perspective, we can say that analytics is a keystone of data science.

From the *disciplinary* perspective, *data science is a new interdisciplinary field* in which to study data and its domain in terms of a data-to-knowledge-to-wisdom thinking for generating data products [11]. Data science integrates *traditionally data-oriented disciplines* such as statistics, informatics and computing with *traditionally data-independent fields* such as communication, management and sociology.

C. The Future of Data Science

It is difficult at this very early stage of data science to predict specific future data science innovation and research, thus the

next-generation data science will need to address the unknown space that is currently invisible to existing science and create new data products. We will need to:

- Deepen our *understanding of data invisibility* (i.e., *invisible data characteristics*) in the hidden and blind spaces (Spaces B and D in Fig. 1 [11]), to understand their X-complexities (see [11]) and X-intelligence (see [11]), and strengthen our capabilities;
- Invent *new data representation capabilities*, including designs, structures, schemas and algorithms to make invisible data in Spaces B and D in Fig. 1 [11] more visible and explicit;
- Create *new analytical and learning capabilities*, including original theories, algorithms and models, to disclose the unknown knowledge in unknown Space D in Fig. 1 [11];
- Build new intelligent systems and services, including corporate and Internet-based collaborative platforms and services, to support collaborative and collective exploration of invisible and unknown challenges in the fully unknown space D in Fig. 1 [11].
- Train a generation of qualified data science professionals in data literacy, thinking, competency, consciousness and cognitive intelligence to work on the above data science agenda.

V. PITFALLS IN DATA SCIENCE

A typical feature of data science being at this very early stage is that different and sometimes contradictory views appear in various communities. It is essential to share and discuss the myths and reality [37], memes [27], and pitfalls to ensure the healthy development of the field. Based on our observations about the relevant communities, and our experience and lessons learned in conducting data science and analytics research, education and services, we list the following myths and pitfalls for discussion.

A. About Data Science Concepts

Typically, data science has been defined in terms of specific disciplinary foundations, principles, goals, inputs, algorithms and models, processes, tools, outputs, applications, and/or professions. Often, a fragmented statement is given, which may cause debate and result in the phenomenon of “how does a blind person recognize an elephant?” In this section, we discuss some common arguments and observations.

- Data science is statistics [4], [26]; “why do we need data science when we’ve had statistics for centuries” [66]? How does data science really differ from statistics [27]? (Comments: Data science provides systematic, holistic and multi-disciplinary solutions for learning explicit and implicit insights and intelligence from complex and large-scale data and generates evidence or indicators from data by undertaking diagnostic, descriptive, predictive and/or prescriptive analytics, in addition to supporting other tasks on data such as computing and management.)
- Why do we need data science when information science and data engineering have been explored for many years?

(Comments: Consider the issues faced in related areas by the enormity of the task and the parallel example of enabling a blind person to recognize an animal as large as an elephant. Information science and data engineering alone cannot achieve this. Other aspects may be learned from the discussion about greater or fewer statistics; more in [16].)

- I have been doing data analysis for dozens of years; data science has nothing new to offer me. (Comments: Classic data analysis and technologies focus mostly on explicit observation analysis and hypothesis testing on small and simpler data.)
- Is data science old wine in a new bottle? What are the new grand challenges foregrounded by data science? (Comments: The analysis of the gaps between existing developments and the potential of data science (see Fig. 2) shows that many opportunities can be found to fill the theoretical gaps when data complexities extend significantly beyond the level that can be handled by the state-of-the-art theories and systems, e.g., classic statistical and analytical theories and systems were not designed to handle the non-IIDness [7] in complex real-life systems.)
- Data science mixes statistics, data engineering and computing, and does not contribute to breakthrough research. (Comments: Data science attracts attention because of the significant complexities in handling complex real-world data, applications and problems that cannot be addressed well by existing statistics, data engineering and computing theories and systems. This drives significant innovation and produces unique opportunities to generate breakthrough theories.)
- Data science is also referred to as data analytics and big data [1]. (Comments: This confuses the main objectives, features, scopes of the three concepts and areas. Data science needs to be clearly distinguished from both data analytics and big data.)
- Other definitions ascribed to data science are that it is big data discovery [22], prediction [24], or the combination of principle and process with technique [51].

It is also worth noting attention that the terms “big data”, “data science” and “advanced analytics” are often extensively misused, over-used or improperly used by diverse communities and for various purposes, particularly given the influence of media hype and buzz. A large proportion of Google searches on these keywords returns results that are irrelevant to their intrinsic semantics and scope, or simply repeat familiar arguments about the needs of data science and existing phenomena. In many such findings [5], [54], [17], [29], [2], [25], [50], [39], [40], [18], [55], [45], [49], [23], [36], [41], [48], [34], [35], [43], [21], [15], [53], [31], [33], big data is described as being simple, data science has nothing to do with the science of data, and advanced analytics is the same as classic data analysis and information processing. There is a lack of deep thinking and exploration of why, what and how these new terms should be defined, developed and applied.

The above observations strongly illustrate that data science is still in a very early stage. They also justify the urgent need to develop sound terminology, standards, a code of conduct, statement and definitions, theoretical frameworks, and better practices that will exemplify typical data science professional practices and profiles.

B. About Data Volume

- What makes data “big”? (Comments: It is usually not the volume but the complexities (as discussed in [11], [8]) and large values that make data big.)
- Why is the bigness of data important? (Comments: The bigness (referring to data science complexities) of data heralds new opportunities for theoretical, technological, practical, economic and other development or revolution.)
- Big data refers to massive volumes of data. (Comments: Here, “big” refers mainly to significant data complexities. From the volume perspective, a data set is big when the size of the data itself becomes a quintessential part of the problem.)
- Data science is big data analytics. (Comments: Data science is a comprehensive field centered on manipulating data complexities and extracting intelligence, in which data can be big or small and analytics is a core component and task.)
- I do not have big data so I cannot do big data research. (Comments: Most researchers and practitioners do not have sizeable amounts of data and do not have access to big infrastructure either. However, significant research opportunities still exist to create fundamentally new theories and tools to address respective X-complexities and X-intelligence.)
- The data I can find is small and too simple to be explored. (Comments: While scale is a critical issue in data science, small data, which is widely available, may still incorporate interesting data complexities that have not been well addressed. Often, we see experimental data, which is usually small, neat and clean. Observational data from real business is live, complex, large and frequently messy.)
- I am collecting data from all sources in order to conduct big data analytics. (Comments: Only relevant data is required to achieve a specific analytical goal.)
- It is better to have too much data than too little. (Comments: While more data generally tends to present more opportunities, the data amount needs to be relevant to the data needed and the data manipulation goals. Whether bigger is better depends on many aspects.)

C. About Data Infrastructure

- I do not have big infrastructure, so I cannot do big data research. (Comments: While big infrastructure is useful or necessary for some big data tasks, theoretical research on significant challenges may not require big infrastructure.)
- My organization will purchase a high performance computer to support big data analytics (Comments: Many

big data analytics tasks can be done without a high performance computer. It is also essential to differentiate between distributed/parallel computing and high performance computing.)

D. About Analytics

- Thinking data-analytically is crucial for data science. (Comments: Data-analytic thinking is not only important for a specific problem-solving, but is essential for obtaining a systematic solution and for a data-rich organization. Converting an organization to think data analytically is a critical competitive advantage in the data era.)
- The task of an analyst is mainly to develop common task frameworks and conduct inference [3] from the particular to the general. (Comments: Analytics in the real world is often specific. Focusing on certain common task frameworks may trigger incomplete or even misleading outcomes. As discussed in Section IV-B, an analyst may take other roles, e.g., predictive modeling is typically problem-specific.)
- I only trust the quality of models built in commercial analytical tools. (Comments: Such tools may produce misleading or even incorrect outcomes if the assumption of their theoretical foundation does not fit the data, e.g., if they only suit imbalanced data, normal distribution-based data, or IID data.)
- Most published models and algorithms and their experimental outcomes are not repeatable. (Comments: Such works seem to be more hand-crafted rather than manufactured. Repeatability, reproducibility, open data and data sharing are critical to the healthy field development.)
- I want to do big data analytics, can you tell me which algorithms and program language I should learn? (Comments: Public survey outcomes (see examples in [9]) give responses to such questions. Which algorithms, language and platform should be chosen also depends on organizational maturity and needs. For long-term purposes, big data analytics is about building competencies rather than specific functions.)
- My organization’s data is private and thus you cannot be involved in our analytics. (Comments: Private data can still be explored by external parties by implementing proper privacy protection and setting up appropriate policies for onsite exploration.)
- Let me (an analyst) show you (business people) some of my findings which are statistically significant. (Comments: As domain-driven data mining [13] shows, many outcomes are often statistically significant but are not actionable. An evaluation of those findings needs to be conducted to discover what business impact [14] might be generated if the findings they may generate are operationalized.)
- Strange, why can I not understand and interpret the outcomes? (Comments: This may be because the problem has been misstated, the model may be invalid for the data, or the data used is not relevant or correct.)

- Your outcomes are too empirical without theoretical proof and foundation. (Comments: While it would be ideal if questions about the outcomes could be addressed from theoretical, optimization and evaluation perspectives, real-life complex data analytics may often be more exploratory and it may initially be difficult to optimize empirical performance.)
- My analysis shows what you delivered is not the best for our organization. (Comments: It may be challenging to claim “the best” when a variety of models, workflows and data features are used in analytics. It is not unusual for analysts to obtain different or contradictory outcomes on the same data as a result of the application of different theories, settings and models. It may turn out to be a very challenging job to find a solid model that perfectly and stably fits the invisible aspect of data characteristics. It is important to appropriately check the relevance and validity of the data, models, frameworks and workflows available and used. Doing the right thing at the right time for the right purpose is a very difficult task when attempting to understand complex real-life data and problems.)
- Can your model address all of my business problems? (Comments: Different models are often are required to address diverse business problems, as a single model cannot handle a problem sufficiently well.)
- This model is very advanced with solid theoretical foundation, let us try it in your business. (Comments: While having solid scientific understanding of a model is important, it is data-driven discovery may better capture the actual data characteristics in real-life problem solving. A model may be improperly used without a deep understanding of model and data suitability. Combining data driven approaches with model driven approaches may be more practical.)
- My analytical reports consist of lots of figures and tables that summarize the data mining outcomes, but my boss seems not so interested in them. (Comments: Analytics is not just about producing meaningful analytical outcomes and reports; rather, it concerns insights, recommendations and communication with upper management for decision-making and action.)
- It is better to have advanced rather than simple models. (Comments: Generally, simpler is better. The key to deploying a model is to fit the model to the data while following the same assumption taken by the model.)
- We just tuned the models last month, but again they do not work well. (Comments: Monitoring a model’s performance by watching the dynamics and significant change that may take place in the data and business is critical. Real-time analytics requires adaptive and automated re-learning and adjustment.)
- I designed the model, so I trust the outcomes. (Comments: The reproducibility of model outcomes relies on many factors. A model that is properly constructed may fall short in other aspects such as data leakage, overfitting, insufficient data cleaning, and poor understanding of data

characteristics and business. Similarly, a lack of communication with the business may cause serious problems in the quality of the outcome.)

- Data science and analytics projects are just other kinds of IT projects. (Comments: While data projects share many similar aspects to mainstream IT projects, certain distinctive features in data, the manipulation process, delivery, and especially the exploratory nature of data science and analytics projects require different strategies, procedures and treatments. Data science projects are more exploratory, ad hoc, decision-oriented and intelligence-driven.)

E. About Capabilities and Roles

- I am a data scientist. (Comments: Lately, it seems that everyone has suddenly become a data scientist. Most data scientists simply conduct normal data engineering and descriptive analytics. Do not expect omnipotence from data scientists.)
- “A human investigative journalist can look at the facts, identify what’s wrong with the situation, uncover the truth, and write a story that places the facts in context. A computer can’t.” [40] (Comments: The success of AlphaGo [32] may show the potential that a data science-enabled computer has to undertake a large proportion of the job a journalist does.)
- My organization wants to do big data analytics, can you recommend some of your PhD graduates to us? (Comments: While data science and advanced analytics tasks usually benefit from the input of PhDs, an organization requires different roles and competencies according to the maturity level of the analytics and the organization.)
- Our data science team consists of a group of data scientists. (Comments: An effective data science team may consist of statisticians, programmers, physicists, artists, social scientists, decision-makers, or even entrepreneurs.)
- A data scientist is a statistical programmer. (Comments: In addition to the core skills of coding and statistics, a data scientist needs to handle many other matters; see discussions in [9].)

F. Other Matters

In addition to the above aspects, there are other matters that require careful consideration in conducting data science and analytics. We list some here.

- Garbage in, garbage out. (Comments: The quality of data determines the quality of output.)
- More complex data, a more advanced model, and better outcomes. (Comments: Good data does not necessarily lead to good outcomes; A good model also does not guarantee good outcomes.)
- More general models, better applicability. (Comments: General models may lead to weaker outcomes on a specific problem. It is not reasonable or practical to expect a single tool for all tasks.)

- More frequent patterns, more interesting. (Comments: It has been shown that frequent patterns mined by existing theories are generally not useful and actionable.)
- We're interested in outcomes, not theories. (Comments: Actionable outcomes may need to satisfy both technical and business significance [13].)
- The goal of analytics is to support decision-making actions, not just to present outcomes about data understanding and analytical results. (Comments: This addresses the need for actionable knowledge delivery [6] to recommend actions from data analytics for decision-support.)
- Whatever you do, you can at least get some values. (Comments: This is true, but it may be risky or misleading. Informed data manipulation and analytics requires a foundation for interpreting why the outcomes look the way they do.)
- Many end users are investing in big data infrastructure without project management. (Comments: Do not rush into data infrastructure investment without a solid strategic plan of your data science initiatives, which requires the identification of business needs and requirements, the definition of reasonable objectives, the specification of timelines, and the allocation of resources.)
- Pushing data science forward without suitable talent. (Comments: On one hand, you should not simply wait for the right candidate to come along, but should, actively plan and specify the skills needed for your organization's initiatives and assemble a team according to the skill-sets required. On the other hand, getting the right people on board is critical, as data science is essentially about intelligence and talent.)
- No culture for converting data science insights into actionable outcomes. (Comments: This may be common in business intelligence and technically focused teams. Fostering a data science-friendly culture requires a top-down approach driven by business needs, and making data-driven decisions that enable data science specialists and project managers to be part of the business process, and to conduct change management.)
- Correct evaluation of outcomes. (Comments: This goes far beyond such technical metrics as Area Under the ROC Curve and Normalized Mutual Information. Business performance after adopting the recommended outcomes needs to be evaluated [6]. For example, recent work on high utility analysis [38] and high impact behavior analysis [14] study how business performance can be taken into data modeling and evaluation account. Solutions that lack business viability are not actionable.)
- Apply a model in a consistent way. (Comments: It is essential to understand the hypothesis behind a model and to apply a model consistent with its hypothesis.)
- Overthinking and overusing models. (Comments: All models and methods are specific to certain hypotheses and scenarios. No models are universal and sufficiently "advanced" to suit everything. Do not assume that if the data is tortured long enough, it will confess to anything.)

- Know nothing about the data before applying a model. (Comments: Data understanding is a must-do step before a model is applied.)
- Analyze data for the sake of analysis only. (Comments: This involves the common bad practice of overusing analytics.)
- What makes an insight (knowledge) actionable? (Comments: This is dependent on not only the statistical and practical values of the insight, but also predictive power and business impact.)
- Do not assume the data you are given is perfect. (Comments: Data quality forms the basis of obtaining good models, outcomes and decisions. Poor quality data, the same as poor quality models, can lead to misleading or damaging decisions. Real-life data often contains imperfect features such as incompleteness, uncertainty, bias, rareness, imbalance and non-IIDness.)

VI. CONCLUSIONS

In the era of data science, big data and advanced analytics, numerous debates have emerged from a wide range of backgrounds, domains, areas and perspectives and for diversified reasons and purposes. It is difficult but critical to explore the nature of data science. To do so, a fundamental perspective is to explore the intrinsic characteristics, challenges, working mechanisms, and dynamics of data and the science about data.

As part of our comprehensive review of data science [10], [9], [11], the discussions about the nature and pitfalls of data science in this work will hopefully stimulate deep and intrinsic discussions about what makes data science a new science, and what makes data science valuable for research, innovation, the economy, services and professionals.

ACKNOWLEDGMENT

This work is partially sponsored by the Australian Research Council Discovery Grant (DP130102691).

REFERENCES

- [1] P. E. Anderson, C. Turner, J. Dierksheide, and R. McCauley. An extensible online environment for teaching data science concepts through gamification. In *2014 IEEE Frontiers in Education Conference (FIE)*, pages 1–8, 2014.
- [2] O. Anya, B. Moore, C. Kieliszewski, P. Maglio, and L. Anderson. Understanding the practice of discovery in enterprise big data science: An agent-based approach. In *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences*, volume 3, pages 882–889, 2015.
- [3] L. Breiman. Statistical modeling: The two cultures. *Statist. Sci.*, 16(3):199–231, 2001.
- [4] K. Broman. Data science is statistics, 2013.
- [5] G. Brown. Review of education in mathematics, data science and quantitative disciplines: Report to the group of eight universities, 2009.
- [6] L. Cao. Domain driven data mining: challenges and prospects. *IEEE Trans. on Knowledge and Data Engineering*, 22(6):755–769, 2010.
- [7] L. Cao. Non-iidness learning in behavioral and social data. *The Computer Journal*, 57(9):1358–1370, 2014.
- [8] L. Cao. *Metasynthetic Computing and Engineering of Complex Systems*. Springer, 2015.
- [9] L. Cao. Data science: A comprehensive overview. *Submitted to ACM Computing Survey*, pages 1–37, 2016.
- [10] L. Cao. Data science and analytics: A new era. *International Journal of Data Science and Analytics*, 1(1):1–2, 2016.

- [11] L. Cao. Data science: Intrinsic challenges and directions, 2016. Technical Report, UTS Advanced Analytics Institute.
- [12] L. Cao, Y. Ou, and P. S. Yu. Coupled behavior analysis with applications. *IEEE Trans. on Knowledge and Data Engineering*, 24(8):1378–1392, 2012.
- [13] L. Cao, P. S. Yu, C. Zhang, and Y. Zhao. *Domain Driven Data Mining*. Springer, 2010.
- [14] L. Cao, Y. Zhao, and C. Zhang. Mining impact-targeted activity patterns in imbalanced data. *IEEE Trans. on Knowledge and Data Engineering*, 20(8):1053–1066, 2008.
- [15] E. Casey. The growing importance of data science in digital investigations. *Digital Investigation*, 14:A1–A2, 2015.
- [16] J. M. Chambers. Greater or lesser statistics: A choice for future research. *Statistics and Computing*, 3(4):182–184, 1993.
- [17] S. Chawla, J. D. Hartline, and D. Nekipelov. Mechanism design for data science. In *Economics and computation: Proceedings of the Fifteenth ACM Conference*, pages 711–712, 2014.
- [18] T. R. Clancy, K. H. Bowles, L. Gelinas, I. Androwich, C. Delaney, S. Matney, J. Sensmeier, J. Warren, J. Welton, and B. Westra. A call to action: Engage in big data science. *Nursing Outlook*, 62(1):64–65, 2014.
- [19] CN. China big data, 2015.
- [20] E. Commission. Commission urges governments to embrace potential of big data. 2014.
- [21] A. Cuzzocrea and M. M. Gaber. Data science and distributed intelligence: Recent developments and future insights. *Studies in Computational Intelligence*, 446:139–147, 2013.
- [22] T. H. Davenport and D. Patil. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, pages 70–76, 2012.
- [23] R. M. de Moraes and L. Martinez. Computational intelligence applications for data science. *Knowledge-Based Systems*, 87:1–2, 2015.
- [24] V. Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.
- [25] H. A. Dierick and F. Gabbiani. Drosophila neurobiology: No escape from ‘big data’ science. *Current Biology*, 25(14):606–608, 2015.
- [26] P. J. Diggle. Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4):793–813, 2015.
- [27] D. Donoho. 50 years of data science, 2015.
- [28] J. H. Faghmous and V. Kumar. A big data guide to understanding climate change: The case for Theory-Guided data science. *Big Data*, 2(3):155–163, 2014.
- [29] J. Faris, E. Kolker, A. Szalay, L. Bradlow, E. Deelman, W. Feng, J. Qiu, D. Russell, E. Stewart, and E. Kolker. Communication and data-intensive science in the beginning of the 21st century. *A Journal of Integrative Biology*, 15(4):213–215, 2011.
- [30] P. Geczy. Big data characteristics. *The Macrotheme Review*, 3(6):94–104, 2014.
- [31] M. Gold, R. McClarren, and C. Gaughan. The lessons oscar taught us: Data science and media & entertainment. *Big Data*, 1(2):105–109, 2013.
- [32] Google. Deepmind, 2016.
- [33] A. Gupta, A. Cecen, S. Goyal, A. K. Singh, and S. R. Kalidindi. Structure-property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system. *Acta Mater*, 91:239–254, 2015.
- [34] D. J. Hand. Statistics and computing: The genesis of data science. *Statistics and Computing*, 25(4):705–711, 2015.
- [35] B. T. Hazena, C. A. Booneb, J. D. Ezellc, and L. A. Jones-Farmer. 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*, 154:72–80, 2014.
- [36] N. J. Horton, B. S. Baumer, and H. Wickham. Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics. *arXiv preprint arXiv:1502.00318*, 2015.
- [37] H. V. Jagadish. Big data and science: Myths and reality. *Big Data Research*, 2(2):49–52, 2015.
- [38] L. C. Junfu Yin, Zhigang Zheng. Uspan: An efficient algorithm for mining high utility sequential patterns. In *KDD 2012*, pages 660–668, 2012.
- [39] J. M. Kanter and K. Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2015.
- [40] K. Kirkpatrick. Putting the data science into journalism. *Communications of the ACM*, 58(5):15–17, 2015.
- [41] M. Loukides. *What is data science?* O’Reilly Media, Sebastopol, CA, 2012.
- [42] A. Manieri, S. Brewer, R. Riestra, Y. Demchenko, M. Hemmje, T. Wiktorski, T. Ferrari, and J. Frey. Data science professional uncovered: How the EDISON project will contribute to a widely accepted profile for data scientists. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 588–593, 2015.
- [43] A. Manieri, F. S. Nucci, M. Femminella, and G. Reali. Teaching Domain-Driven data science: Public-Private co-creation of Market-Driven certificate. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 569–574, 2015.
- [44] McKinsey. Big data: The next frontier for innovation, competition, and productivity, 2011. McKinsey Global Institute.
- [45] C. C. Miller. Data science: The numbers of our lives. *New York Times*, 2013.
- [46] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2011.
- [47] NCSU. Institute for advanced analytics, north carolina state university, 2007.
- [48] C. O’Neil and R. Schutt. *Doing data science: Straight talk from the frontline*. O’Reilly Media, Sebastopol, CA, 2013.
- [49] S. K. Pal, S. K. Meher, and A. Skowron. Data science, big data and granular mining. *Pattern Recognition Letters*, 67(2):109–112, 2015.
- [50] T. Priebe and S. Markus. Business information modeling: A methodology for data-intensive projects, data science and big data governance. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2056–2065, 2015.
- [51] F. Provost and T. Fawcett. Data science and its relationship to big data and Data-Driven decision making. *Big Data*, 1(1):51–59, 2013.
- [52] C. Rudin. *Discovery with data: Leveraging statistics with computer science to transform science and society*, 2014. American Statistical Association.
- [53] C. Siart, S. Kopp, and J. Apel. The interface between data science, research assessment and science support - highlights from the german perspective and examples from heidelberg university. In *2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 472–476, 2015.
- [54] J. Stanton. An introduction to data science, 2012.
- [55] M. L. Stevens. An ethically ambitious higher education data science. *Research & Practice in Assessment*, 9:96–97, 2014.
- [56] M. Stonebraker, S. Madden, and P. Dubey. Intel ‘big data’ science and technology center vision and execution plan. *SIGMOD Record*, 42(1):44–49, 2013.
- [57] J. W. Tukey. The future of data analysis. *Ann. Math. Statist.*, 33(1):1–67, 1962.
- [58] J. W. Tukey. *Exploratory Data Analysis*. Pearson, 1977.
- [59] UK. UK big data, 2016.
- [60] UN. United nation global pulse projects, 2010.
- [61] USNSF. US big data research initiative, 2012.
- [62] UTSAAI. Advanced analytics institute, university of technology sydney, 2011.
- [63] D. van Dyk, M. Fuentes, M. I. Jordan, M. Newton, B. K. Ray, D. T. Lang, and H. Wickham. ASA statement on the role of statistics in data science, 2015.
- [64] M. A. Walker. The professionalisation of data science. *Int. J. of Data Science*, 1(1):7–16, 2015.
- [65] WEF. The global competitiveness report 2011–2012: An initiative of the world economic forum, 2011.
- [66] I. Wladawsky-Berger. Why do we need data science when we’ve had statistics for centuries? *The Wall Street Journal*, 2014.
- [67] J. Wu. Statistics = data science?, 1997.