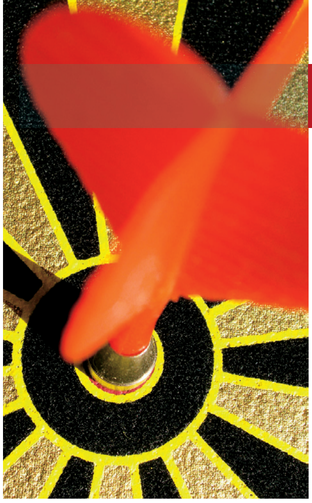

EXPERT OPINION

Editor: **Daniel Zeng**, University of Arizona and Chinese Academy of Sciences, zengdaniel@gmail.com

Data Science: Nature and Pitfalls

Longbing Cao, *University of Technology, Sydney*

The era of analytics,[1] data science,[2] and big data[3] has driven substantial governmental, industrial, and disciplinary interest; goal and strategy transformation; and a paradigm shift in research and innovation. This has resulted in significant opportunities and prospects becoming available, and an overwhelming amount of fanfare has spread across domains, areas, and events.

A review of the related initiatives, progress, and status of data science, analytics, and big data[4] and the diversified discussions about the prospects, challenges, and directions[5] makes clear the controversy caused by the potential conflict of these various elements. There is a need for deep discussions about the nature and pitfalls of data science, clarification of fundamental concepts and myths, and a demonstration of the intrinsic characteristics and opportunities of data science.

Thus, this article focuses on two fundamental issues—the nature and pitfalls of data science. I highlight the status, intrinsic factors, characteristics, and features of the era of data science and analytics, as well as the challenges and opportunities in innovation, research, and disciplinary development. I also summarize common pitfalls about the concepts of data science, data volume, infrastructure, analytics, and capabilities and roles. Building on these discussions, I then present the concepts and possible future directions of data science.

Features of the Data Science Era

Identifying the features and characteristics of the data science era is critical and challenging. Let's explore this from the perspective of the transformation and paradigm shift caused by data science and discuss the core driving forces and the status of several typical issues confronting the data science field.

Transformation and Paradigm Shift

The emergence of the era of data science and analytics can be highlighted by three key indicators:

- a disciplinary paradigm shift, or the shifting of data-centric disciplinary paradigms from one to another;
- technological transformation, or the upgrading of data technology from one generation to another; and
- innovative production, or the innovation of technical and practical data products.

We can define the disciplinary paradigm shift of data-oriented and data-centric research, innovation, and professions as moving from data analysis to data analytics, from descriptive analytics to deep analytics, and from data analytics to data science. The disciplinary paradigm shift promotes data-related technological transformation from large-scale data to big data, from business operational systems to business analytical systems, from the World Wide Web to the Wisdom Web, and from the Internet to the Internet of Everything (including mobile and social networks and the Internet of Things).

Innovative production in data and analytics can be represented by typical indicators—for example, from a digital to a data economy, from closed to open government, from e-commerce to online business, from landlines to smartphones, and from

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

the Internet to mobile and social networks.

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 the following:

- data, which involves historical, real-time, and future data;
- behavior, which bridges the gaps between the physical and data worlds;
- complexity, which differentiates one data system from another;
- intelligence, which is embedded in a data system;
- service, which is present in various forms and domains; and
- opportunities, because data enables enormous opportunities.

Data-enabled opportunities, also called *X-opportunities*, are overwhelming. They extend from research, innovation, education, government, and the economy, and can include opportunities in the following areas:

- research, such as inventing data-focused breakthrough theories and technologies;
- innovation, such as developing cutting-edge data-based services, systems, and tools;
- education, such as innovating data-oriented courses and training;
- government, such as enabling data-driven government decision making and objectives;
- economics, such as fostering data economy, services, and industrialization;
- lifestyle, such as promoting data-enabled smarter living and smarter cities; and
- entertainment, such as creating data-driven entertainment activities, networks, and societies.

A data science problem is a complex system[6,7] in which comprehensive system complexities, also called *X-complexities*,[5] are embedded. These comprise complexities in the following areas:

- data, including comprehensive data circumstances and characteristics;
- behavior, including individual and group activities, evolution, utility, impact, and change;
- the domain, including domain factors, processes, norms, policies, knowledge, and domain expert engagement in problem solving;
- social complexity, including social networking, community formation and divergence, sentiment, the dissemination of opinion and influence, and other social issues such as trust and security;
- the environment, including contextual factors, interactions with systems, changes, and uncertainty;
- learning, including the development of appropriate methodologies, frameworks, processes, models and algorithms, and theoretical foundations and explanations; and
- decision making, including the methods and forms of deliverables, communications, and decision-making actions.

In a complex data science problem, ubiquitous intelligence, also called *X-intelligence*,[5] is often demonstrated and must be incorporated and synergized[7] in problem-solving processes and systems:

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

All of these 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.

Data DNA

As a result of data quantification, data is everywhere, including the Internet; the IoT; sensor networks; sociocultural, 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* 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, instance, human, organization, system, or part of a subsystem. Property refers to the attributes that describe an entity. Relationship corresponds to entity interactions and 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 often comprises many diverse entities, each of which has specific properties, and different relationships are embedded within and between properties and entities. From the lowest to the highest levels, 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 supporting personalized characteristics 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's 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[8] and nonpersonal 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 Wi-Fi usage data. Such student data DNA is both fixed and evolving.

In complex data, data DNA is embedded within various X-complexities and ubiquitous X-intelligence in a data organism.[5,7] This makes data rich in content, characteristics, semantics, and value, but challenging in acquisition, preparation, presentation, analysis, and interpretation.

Data Quality

Data science tasks involve roles and follow processes that differ from more generalized IT projects, because data science and analytics works tend to be creative, intelligent, exploratory, nonstandard, unautomated, and personalized, and they 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 must be considered 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, or 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.

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 must 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 data's dynamic nature. Because of the changing nature of data, the need to check the validity, veracity, and reliability data used and analytics undertaken is very 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.

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 in data, modeling, and deliverables.

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 both destructive forces and unauthorized users' actions, such as improper use or disclosure. It addresses not only privacy issues but 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, traceability, liability, and warranty of both the measurement and results, as well as the efficacy and verifiability of the analytics and protection.

Data and analytics trust refers to the belief in the reliability, truth, or ability of data and analytics to achieve 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.

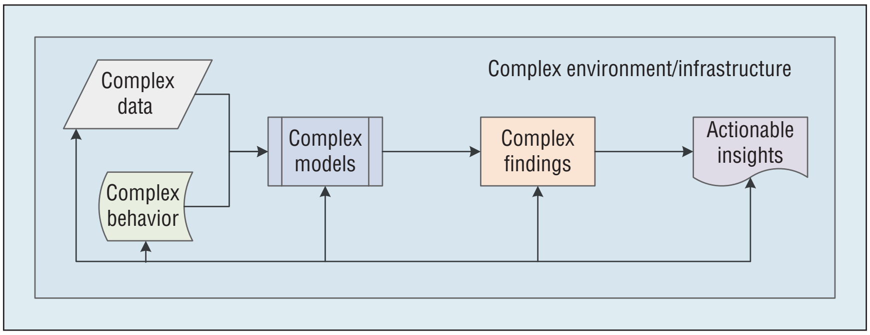

Figure 1. The extreme data challenge.

The Extreme Challenge

Different types and levels of analytical problems trouble the existing knowledge base, and we are especially challenged by the problems in complex data and environments. Our focus on data science research and innovation concerns what we call an *extreme data challenge* in data science and analytics. The extreme data challenge illustrated in Figure 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 focus on the following:

- complex data with complex characteristics[5,7];
- complex behaviors with complex relationships and dynamics[5,7];
- complex environments in which complex data and behaviors are embedded and interacted with[5,7];
- complex models to address the data and behavior complexities in a complex environment;
- complex findings to uncover hidden but technically interesting and business-friendly observations, indicators or evidence, statements, or presentations; and
- actionable insights to demonstrate the next best or worst situation and inform the optimal strategies to support effective business decision making.[9,10]

Many real-life problems fall into this level of complexities and challenges, as the extreme data challenge shows, and they have not been addressed well. One example is understanding group behaviors by multiple actors when there are complex interactions and relationships, such as in the manipulation of large-scale cross-capital markets pool by internationally collaborative investors,[11] each of whom plays a role by connecting information from the underlying markets, social media, other financial markets, socioeconomic data, and policies.[12] Another example would be to predict local climate change and effect by connecting local, regional, and global climate, geographical, and agricultural data and other information.[13]

Disciplinary Development of Data Science

I present a status summary of the disciplinary development of data science by reviewing the developmental gaps between the data's potential and the state-of-the-art capabilities to fulfill such potential, the research map of data science, and the course framework of data science.

Data-to-Capability Development Gaps

The rapid increase in big data has led to significant gaps between what is in

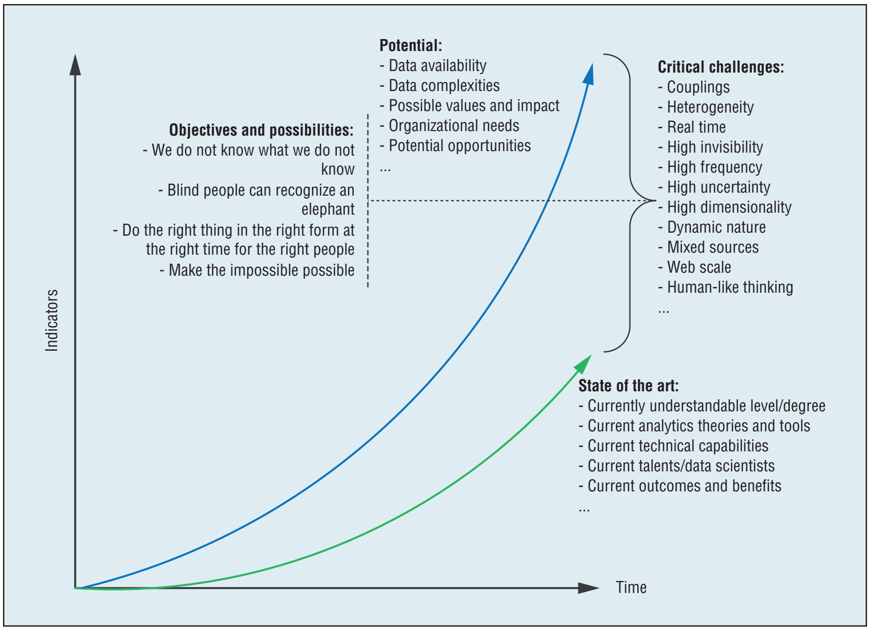
Potential:
- Data availability
- Data complexities
- Possible values and impact
- Organizational needs
- Potential opportunities
...
Critical challenges:
- Couplings
- Heterogeneity
- Real time
- High invisibility
- High frequency
- High uncertainty
- High dimensionality
- Dynamic nature
- Mixed sources
- Web scale
- Human-like thinking
...
Objectives and possibilities:
- We do not know what we do not know
- Blind people can recognize an elephant
- Do the right thing in the right form at the right time for the right people
- Make the impossible possible
Indicators
State of the art:
- Currently understandable level/degree
- Current analytics theories and tools
- Current technical capabilities
- Current talents/data scientists
- Current outcomes and benefits
...
Time

Figure 2. Critical development gaps between data potential and state-of-the-art capabilities.

the data and how much we can understand.[9,10] Figure 2 shows empirically the data development gaps between the growth of data potentials and the state-of-the-art capabilities. Such gaps have increased in the past 10 years, especially recently, owing to the imbalance between potential exponential increase and progressive state-of-the-art capability development. Examples of such gaps could include the gaps between

- data availability and the currently understandable data level, scale, and degree;
- data complexities and the currently available analytics theories and tools;
- data complexities and the currently available technical capabilities;
- possible values and impact and currently achievable outcomes and benefits;
- organizational needs and the currently available talent (that is, data scientists); and
- potential opportunities and the 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 data that is not independent and identically distributed (IID),[12] which cannot be simplified in such a way that it 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 high-frequency data in the "11-11" shopping seasons launched by Alibaba, or identifying suspects in an imbalanced and multisource data and environment such as fraud detection in high-frequency market trading. Other challenges are high invisibility, high frequency, high uncertainty, high dimensionality, the dynamic nature, mixed sources, online learning at the Web scale, and the development of human-like thinking.

Data Science Research Map

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 requires understanding the problem's intrinsic nature, characteristics, complexities, and boundaries and then analyzing the gaps between the problem complexities and the existing capability set. This gap analysis is critical for original research and breakthroughs in scientific discovery.
- Goal-driven discovery requires understanding the business, technical, and decision goals to be achieved by understanding the problem and then conducting gap analysis of what has been implemented and achieved and what is expected to be achieved.
- Data-driven discovery requires understanding the data characteristics, complexities, and challenges and the gaps between the nature of a problem and the data capabilities. Because of 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" can be biased, dishonest, or manipulated. As a result, the data does not completely capture the problem and thus cannot create a full picture of it through any type of data exploration.

There are two ways to explore major research challenges: to summarize what concerns the relevant communities, and to scrutinize the potential issues arising from the intrinsic complexities and nature of data science problems as complex systems.[7] With the first approach, we can grasp the main research challenges by summarizing the main topics and issues in the statistics communities,[14] informatics and computing communities,[5,15] vendors,[16] government initiatives,[17,18] and research institutions[19,20] that focus on data science and analytics. The second approach is much more challenging, because we must explore the unknown space of the complexities and comprehensive intelligence in complex data problems.

We list some of the data science community's main challenges in addressing the big data complexities represented by key topics in the Data A–Z dictionary. We categorize these as challenges in data/business understanding; mathematical and statistical foundations; X-analytics and data/knowledge engineering; data quality and social issues; data value, impact, and usability; and data to decision and actions.

X-analytics and data/knowledge engineering encompass many specific research issues that have not been addressed properly. These include 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; and networking, communication, and interoperation.

Data Science Course Framework

Data science and analytics education aims 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[21,22] 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's or PhD program in data science.

Public courses are designed for the general community to lift their understanding, skills, profession, and specialism in data science through multilevel short courses. They range from basic courses to intermediate 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 the entire enterprise's innovation and productivity. This involves offering courses and workshops for the workforce, from senior corporate executives to business owners, business analysts, data modelers, data scientists, data engineers, and deployment and enterprise strategists. Such courses cover topics such as data science, data engineering, analytics science, decision science, data and analytics software engineering, project management, communications, and case management.

Undergraduate courses can be offered on either a general data science basis, focusing 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 might be offered to train professionals who will gain knowledge and abilities across disciplines such as business and analytics or statistics and computing.

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

A PhD in data science and analytics program 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.

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

Data Science as a New Science

So, what makes data science a new science? To address this question, we discuss Data A–Z, which can be used to capture every aspect of data science to form a data science ontological system built on discussions about the features, disciplinary development, and future of data science.

Data A–Z

The big data community often uses multiple "V"s to describe the characteristics, challenges, and opportunities of big data. These include volume (size), velocity (speed), variety (diversity), veracity (quality and trust), value (insight), visualization, and variability (formality).

In fact, these terms cannot completely describe big data or the field of data science. Therefore, it is valuable to build a Data A–Z dictionary to capture the intrinsic comprehensive but

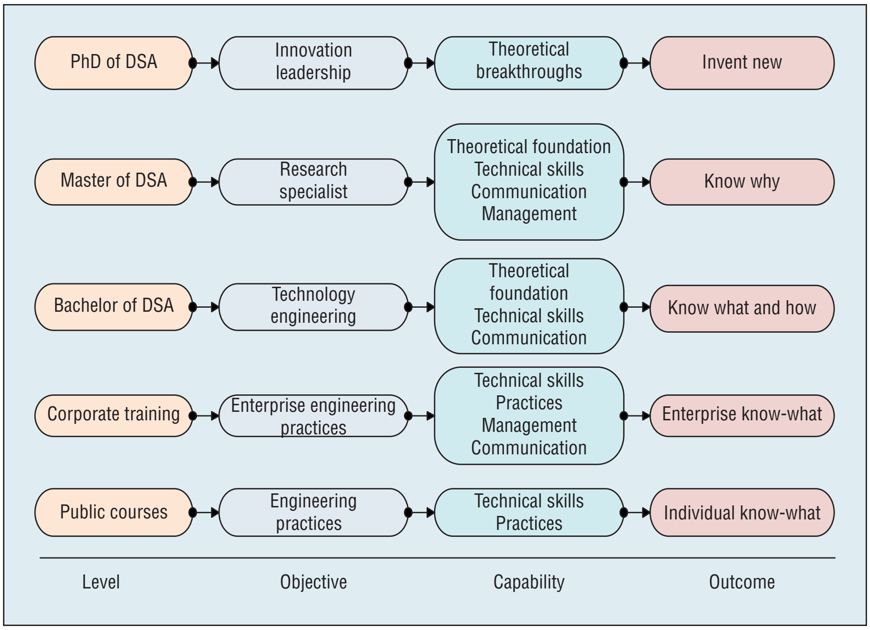

Figure 3. Data science course framework.

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; Zettabyte.

Figure 4. Sample sequence of data science keywords from the Data A–Z dictionary.

diverse aspects, characteristics, challenges, domains, tasks, processes, purposes, applications, and outcomes of data. Figure 4 lists a sample sequence of data science keywords.

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.[23]

What Is Data Science?

Generally speaking, data science is the science (or study) of data. We can define data science as being object focused, process based, or discipline oriented.[5]

> **Definition 2.** From the process perspective, *data science* is a systematic approach to "thinking with wisdom," "understanding domain," "managing data," "computing with data," "mining on knowledge," "communicating with stakeholders," "delivering products," and "acting on insights."

In contrast, data analytics understands data and its underlying business, discovers knowledge, delivers actionable insights, and enables 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.[5] Data science integrates traditionally data-oriented disciplines such as statistics, informatics, and computing with traditionally data-independent fields such as communication, management, and sociology.

The Future of Data Science

It is difficult at this early stage of data science to predict specific future data science innovation and research; thus, next-generation data science will need to address the unknown space that is invisible to existing science and create new data products. We will need to deepen our understanding of data invisibility (that is, invisible data characteristics) in the hidden and blind spaces, understand their X-complexities and X-intelligences, and strengthen our capabilities.[5] We will need to invent data representation capabilities, including designs, structures, schemas, and algorithms, to make invisible data more visible and explicit. Another task will be creating analytical and learning capabilities, including original theories, algorithms, and models, to disclose the unknown knowledge in unknown spaces.[5] New intelligent systems and services will need to be built, including corporate and Internet-based collaborative platforms and services, to support collaborative and collective exploration of invisible and unknown challenges in fully unknown spaces.[5] Finally, we will need to train a generation of qualified data science professionals in data literacy, thinking, competency, consciousness, and cognitive intelligence to work on this data science agenda.

Data Science Pitfalls

At this early stage, it is typical to see different and sometimes contradictory views appear in various communities. It is essential to share and discuss the myths and reality,[24] memes,[25] and pitfalls to ensure the healthy development of the field. From observations about the relevant communities, as well as experiences and lessons learned in conducting data science and analytics research, education, and services, several myths and pitfalls have emerged.

Pitfalls 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 professions. Often, a fragmented statement can cause debate and result in the phenomenon of "How does a blind person recognize an elephant?" Here, we discuss some common arguments and observations.

Data science is statistics,[26,27] so one might argue, "Why do we need data science when we've had statistics for centuries?"[28] or "How does data science really differ from statistics?"[25] Data science provides systematic, holistic, and multidisciplinary 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 prescriptive analytics, in addition to supporting other tasks on data, such as computing and management

Others might ask, "Why do we need data science when information science and data engineering have been explored for many years?" But 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.[14]

Another objection could come from those who have been doing data analysis for decades and who believe that data science has nothing new to offer them. We would counter that classic data analysis and technologies focus mostly on explicit observation analysis and hypothesis testing on small and simpler data.

Others might wonder whether data science is old wine in a new bottle, or what new grand challenges are foregrounded by data science. Analysis of the gaps between existing developments and data science's potential (see Figure 2) shows that many opportunities exist to fill the theoretical gaps when data complexities extend significantly beyond the level that can be handled by state-of-the-art theories and systems. For example, classic statistical and analytical theories and systems were not designed to handle non-IIDness[12] in complex real-life systems.

It is also worth noting attention that the terms *big data*, *data science*, and *advanced analytics* are often overly used or improperly used. Most Google searches on these keywords return 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, 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.

These observations illustrate that data science is still young. 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.

Data Volume Pitfalls

Various pitfalls surround data volume. For example, what makes data "big"? It is usually not the volume but the complexities[5,7] and large values that make data big. Why is the bigness of data important? The bigness (referring to data science complexities) of data heralds new opportunities for theoretical, technological, practical, and economic developments.

Some assume that big data refers to massive volumes of data, but actually, it refers mainly to significant data complexities. From the volume perspective, a dataset is big when the size of the data itself becomes a quintessential part of the problem.

One might argue, "I do not have big data, so I cannot do big data research." 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.

Another idea we see is to collect data from all sources in order to conduct big data analytics. Actually, only relevant data is required to achieve a specific analytical goal. A related pitfall is the idea that it is better to have too much data than too little. Although 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.

Data Infrastructure Pitfalls

Two pitfalls are related to data infrastructure. First, researchers who do not have big infrastructure might think they cannot do big data research. Although big infrastructure is useful or necessary for some big data tasks, theoretical research on significant challenges might not require big infrastructure. Second, an organization will purchase a high-performance computer to support big data analytics, when many big data analytics tasks can be done without one. We must also differentiate between distributed and parallel computing and HPC.

Analytics Pitfalls

There are many pitfalls relating to analytics. For example, consider the idea that data-analytical thinking is crucial for data science. Actually, it is not only important for solving a specific problem, it is essential for obtaining a systematic solution. Converting an organization to think data analytically is a critical competitive advantage in the data era.

Analysts might think that their task is to develop common task frameworks and conduct inference[29] from the particular to the general. In the real world, analytics is often specific. Focusing on certain common task frameworks could trigger incomplete or even misleading outcomes. As we discussed earlier, an analyst can take on other roles; predictive modeling is typically problem specific.

Some researchers might trust the quality of models built in commercial analytical tools alone. However, such tools can produce misleading or even incorrect outcomes if the assumption of their theoretical foundation does not fit the data.

Analysts often want to show "business people" some of their statistically significant findings. But as domain-driven data mining shows,[10] many outcomes are often statistically significant but are not actionable. An evaluation of those findings needs to be conducted to discover what business impact[30] might be generated if the findings they generate are operationalized. Analysts may also wonder why they cannot understand and interpret the outcomes. This could be because the problem has been misstated, the model is invalid for the data, or the data used is not relevant or correct.

Another pitfall is when analytical reports comprise many figures and tables that summarize the data mining outcomes, but the boss does not seem interested in them. 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.

Finally, there is the argument that data science and analytics projects are just other kinds of IT projects. Although 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.

Pitfalls about Capabilities and Roles

Other pitfalls concern the individual—for example, "I am a data scientist." 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.

When an organization wants to do big data analytics, it might seek recommendations of PhD graduates. Although 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.

Another organization might boast that its data science team comprises a group of data scientists. However, an effective data science team could consist of statisticians, programmers, physicists, artists, social scientists, decision makers, or even entrepreneurs.

Finally, there is the argument that a data scientist is a statistical programmer. Although this is true, in addition to the core skills of coding and statistics, a data scientist needs to handle many other matters.[4]

Other Matters

Some additional matters require careful consideration in conducting data science and analytics. I list a few of the common remarks, with my comments in parentheses:

- Garbage in, garbage out. (The quality of data determines the quality of output.)
- More complex data and a more advanced model lead to better outcomes. (Good data does not necessarily lead to good outcomes, nor does a good model.)
- Analytics aims to support decision-making actions, not just to present outcomes about data understanding and analytical results. (This addresses the need for actionable knowledge delivery[9] to recommend actions from data analytics for decision support.)
- Many end users are investing in big data infrastructure without project management. (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. (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 skillsets required. On the other hand, getting the right people on board is critical, because data science is essentially about intelligence and talent.)
- Know nothing about the data before applying a model. (Understanding the data is a must-do step before a model is applied.)
- Do not assume the data you are given is perfect. (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.)

This is just a partial list of the remaining pitfalls in conducting data science and analytics.

As part of this comprehensive review of data science,[4,5,31] I hope that the discussions about the nature and pitfalls of data science in this article will stimulate deep and intrinsic discussions about what makes data science a new science and what makes it valuable for research, innovation, the economy, services, and professionals.

Acknowledgments

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

References

1. J.W. Tukey, "The Future of Data Analysis," *Annals of Mathematical Statistics*, vol. 33, no. 1, 1962, pp. 1–67.
2. J.W. Tukey, *Exploratory Data Analysis*, Pearson, 1977.
3. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Inst., 2011.
4. L. Cao, *Data Science: A Comprehensive Overview*, tech. report, UTS Advanced Analytics Inst., 2016.
5. L. Cao, *Data Science: Intrinsic Challenges and Directions*, tech. report, UTS Advanced Analytics Inst., 2016.
6. M. Mitchell, *Complexity: A Guided Tour*, Oxford Univ. Press, 2011.
7. L. Cao, *Metasynthetic Computing and Engineering of Complex Systems*, Springer, 2015.
8. K. Schwab, *The Global Competitiveness Report 2011–2012*, report, World Economic Forum, 2011.
9. L. Cao, "Domain Driven Data Mining: Challenges and Prospects," *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 6, 2010, pp. 755–769.
10. L. Cao et al., *Domain Driven Data Mining*, Springer, 2010.
11. L. Cao, Y. Ou, and P.S. Yu, "Coupled Behavior Analysis with Applications," *IEEE Trans. Knowledge and Data Eng.*, vol. 24, no. 8, 2012, pp. 1378–1392.
12. L. Cao, "Non-IIDness Learning in Behavioral and Social Data," *Computer J.*, vol. 57, no. 9, 2014, pp. 1358–1370.
13. J.H. Faghmous and V. Kumar, "A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science," *Big Data*, Sept. 2014, pp. 155–163.
14. J.M. Chambers, "Greater or Lesser Statistics: A Choice for Future Research," *Statistics and Computing*, vol. 3, no. 4, 1993, pp. 182–184.
15. C. Rudin, *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society*, Am. Statistical Assoc., 2014.
16. M. Stonebraker, S. Madden, and P. Dubey, "Intel 'Big Data' Science and Technology Center Vision and Execution Plan," *SIGMOD Record*, vol. 42, no. 1, 2013, pp. 44–49.
17. *US Big Data Research Initiative*, Nat'l Science Foundation, 2012.
18. *United Nation Global Pulse Projects*, 2010; www.unglobalpulse.org/projects.
19. Advanced Analytics Inst., Univ. of Technology Sydney, 2011; www.uts.edu.au/research-and-teaching/our-research/advanced-analytics-institute.
20. Inst. for Advanced Analytics, North Carolina State Univ., 2007; http://analytics.ncsu.edu.
21. M.A. Walker, "The Professionalization of Data Science," *Int'l J. Data Science*, vol. 1, no. 1, 2015, pp. 7–16.
22. A. Manieri et al., "Data Science Professional Uncovered: How the EDISON Project Will Contribute to a Widely Accepted Profile for Data Scientists," *Proc. IEEE 7th Int'l Conf. Cloud Computing Technology and Science*, 2015, pp. 588–593.
23. P. Geczy, "Big Data Characteristics," *Macrotheme Rev.*, vol. 3, no. 6, 2014, pp. 94–104.
24. H.V. Jagadish, "Big Data and Science: Myths and Reality," *Big Data Research*, vol. 2, no. 2, 2015, pp. 49–52.
25. D. Donoho, "50 Years of Data Science," 2015; http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.
26. K. Broman, "Data Science Is Statistics," blog, 2013; http://kbroman.wordpress.com/2013/04/05/data-science-is-statistics.
27. P.J. Diggle, "Statistics: A Data Science for the 21st Century," *J. Royal Statistical Society: Series A*, vol. 178, no. 4, 2015, pp. 793–813.
28. I. Wladawsky-Berger, "Why Do We Need Data Science When We've Had Statistics for Centuries?" *Wall Street J.*, 2014; http://on.wsj.com/1iRCO1w.
29. L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science*, vol. 16, no. 3, 2001, pp. 199–231.
30. L. Cao, Y. Zhao, and C. Zhang, "Mining Impact-Targeted Activity Patterns in Imbalanced Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 8, 2008, pp. 1053–1066.
31. L. Cao, *Data Science: Profession and Education*, tech. report, UTS Advanced Analytics Inst., 2016.

Longbing Cao is a professor at the Advanced Analytics Institute at the University of Technology, Sydney. Contact him at longbing.cao@gmail.com.