DOI:10.1145/3015456

While it may not be possible to build a data brain identical to a human, data science can still aspire to imaginative machine thinking.

BY LONGBING CAO

Data Science: Challenges and Directions

WHILE DATA SCIENCE has emerged as an ambitious new scientific field, related debates and discussions have sought to address why science in general needs data science and what even makes data science a science. However, few such discussions concern the intrinsic complexities and intelligence in data science problems and the gaps in and opportunities for data science research. Following a comprehensive literature review,[5,6,10–12,15,18] I offer a number of observations concerning big data and the data science debate. For example, discussion has covered not only data-related disciplines and domains like statistics, computing, and informatics but traditionally less data-related fields and areas like social science and business management as well. Data science has thus emerged as a new inter- and cross-disciplinary field. Although many publications are available, most (likely over 95%) concern existing concepts and topics in statistics, data mining, machine learning, and broad data analytics. This limited view demonstrates how data science has emerged from existing core disciplines, particularly statistics, computing, and informatics. The abuse, misuse, and overuse of the term "data science" is ubiquitous, contributing to the hype, and myths and pitfalls are common.[4] While specific challenges have been covered,[13,16] few scholars have addressed the low-level complexities and problematic nature of data science or contributed deep insight about the intrinsic challenges, directions, and opportunities of data science as an emerging field.

Data science promises new opportunities for scientific research, addressing, say, "What can I do now but could not do before, as when processing large-scale data?"; "What did I do before that does not work now, as in methods that view

» key insights

- **Data science problems require systematic thinking, methodologies, and approaches to help spur development of machine intelligence.**
- **The conceptual landscape of data science assists data scientists trying to understand, represent, and synthesize the complexities and intelligence in related problems.**
- **Data scientists aim to invent data- and intelligence-driven machines to represent, learn, simulate, reinforce, and transfer human-like intuition, imagination, curiosity, and creative thinking through human-data interaction and cooperation.**

data objects as independent and identically distributed variables (IID)?"; "What problems not solved well previously are becoming even more complex, as when quantifying complex behavioral data?''; and "What could I not do better before, as in deep analytics and learning?"

As data science focuses on a systematic understanding of complex data and related business problems,[5,6] I take the view here that data science problems are complex systems[3,19] and data science aims to translate data into insight and intelligence for decision making. Accordingly, I focus on the complexities and intelligence hidden in complex data science problems, along with the research issues and methodologies needed to develop data science from a complex-system perspective.

What It Is

The concept of data science was originally proposed within the statistics and mathematics community[23,24] where it essentially concerned data analysis. Data science today[17] goes beyond specific areas like data mining and machine learning or whether it is the next generation of statistics.[9,11,12] But what is it?

Definition. Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology, to study data following "data science thinking"[6] (see Figure 1). Consider this discipline-based data science formula

data science = {statistics ∩ informatics ∩ computing ∩ communication ∩ sociology ∩ management | data ∩ domain ∩ thinking }

where "|" means "conditional on."

X-Complexities in Data Science

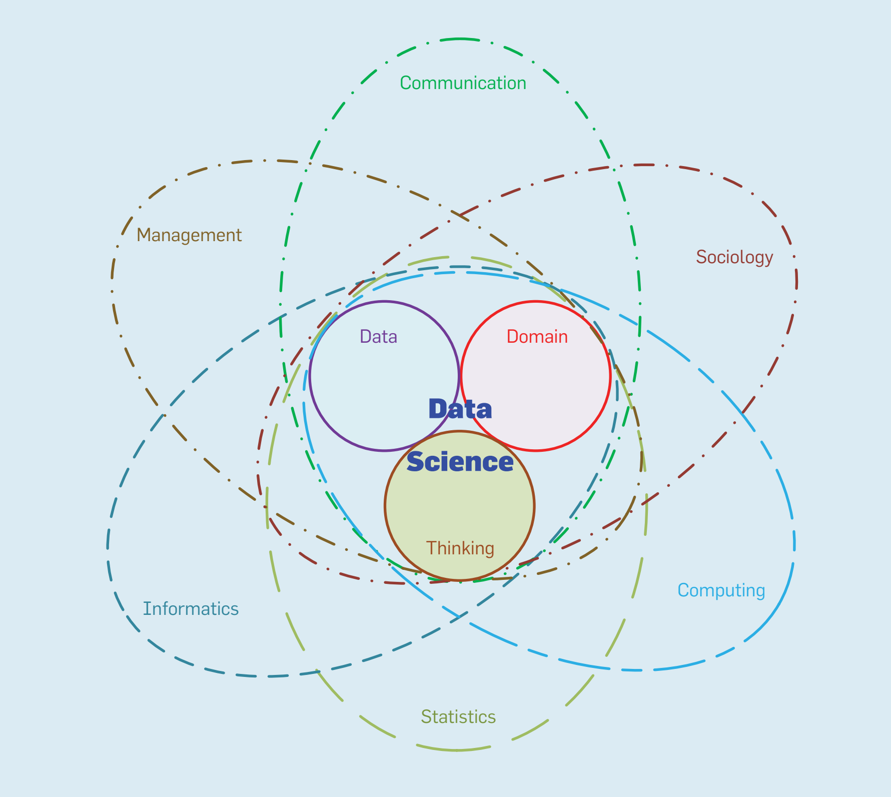
A core objective of data science is exploration of the complexities[19] inherently trapped in data, business, and problem-solving systems.[3] Here, complexity refers to sophisticated characteristics in data science systems. I treat data science problems as complex systems involving comprehensive system complexities, or X-complexities, in terms of data (characteristics), behavior, domain, social factors, environment (context), learning (process and system), and deliverables.

Data complexity is reflected in terms of sophisticated data circumstances and characteristics, including large scale, high dimensionality, extreme imbalance, online and real-time interaction and processing, cross-media applications, mixed sources, strong dynamics, high frequency, uncertainty, noise mixed with data, unclear structures, unclear hierarchy, heterogeneous or unclear distribution, strong sparsity, and unclear availability of specific sometimes critical data. An important issue for data scientists involves the complex relations hidden in data that are critical to understanding the hidden forces in data. Complex relations could consist of comprehensive couplings[2] that may not be describable through existing association, correlation, dependence, and causality theories and systems. Such couplings may include explicit and implicit, structural and nonstructural, semantic and syntactic, hierarchical and vertical, local and global, traditional and nontraditional relations, and evolution and effect.

Data complexities inspire new perspectives that could not have been done or done better before. For example, traditional large surveys of sensor data, including statisticians' questions and survey participants, have been shown to be less effective, as seen in related complications (such as wrongly targeted participants, low overall response rate, and questions unanswered). However, data-driven discovery can help determine who is to be surveyed, what questions need to be answered, the actionable survey operation model, and how cost-effective the survey would be.

Behavior complexity refers to the challenges involved in understanding what actually takes place in business activities by connecting to the semantics and processes and behavioral subjects and objects in the physical world often ignored or simplified in the data world generated by physical-activity-to-data conversion in data-acquisition and -management systems. Behavior complexities are embodied in coupled individual and group behaviors, behavior networking, collective behaviors, behavior divergence and convergence, "nonoccurring"[8] behaviors, behavior-network evolution, group-behavior reasoning, recovery of what actually happened, happens, or will happen in

Figure 1. Transdisciplinary data science.



the physical world from the highly deformed information collected in the purely data world, insights, impact, utility, and effect of behaviors, and the emergence and management of behavior intelligence. However, limited systematic research outcomes are available for comprehensively quantifying, representing, analyzing, reasoning about, and managing complex behaviors.

Data scientists increasingly recognize domain complexity[7] as a critical aspect of data science for discovering intrinsic data characteristics, value, and actionable insight. Domain complexities are reflected in a problem domain as domain factors, domain processes, norms, policies, qualitative-versus-quantitative domain knowledge, expert knowledge, hypotheses, meta-knowledge, involvement of and interaction with domain experts and professionals, multiple and cross-domain interactions, experience acquisition, human-machine synthesis, and roles and leadership in the domain. However, existing data analytics focuses mainly on domain knowledge.

Social complexity is embedded in business activity and its related data and is a key part of data and business understanding. It may be embodied in such aspects of business problems as social networking, community emergence, social dynamics, impact evolution, social conventions, social contexts, social cognition, social intelligence, social media, group formation and evolution, group interaction and collaboration, economic and cultural factors, social norms, emotion, sentiment and opinion influence processes, and social issues, including security, privacy, trust, risk, and accountability in social contexts. Promising interdisciplinary opportunities emerge when social science meets data science.

Environment complexity is another important factor in understanding complex data and business problems, as reflected in environmental (contextual) factors, contexts of problems and data, context dynamics, adaptive engagement of contexts, complex contextual interactions between the business environment and data systems, significant changes in business environment and their effect on data systems, and variations and uncertainty in interactions between business data and the business environment. Such aspects of the system environment have concerned open complex systems[20] but not yet data science. If ignored, a model suitable for one domain might produce misleading outcomes in another, as is often seen in recommender systems.

Learning (process and system) complexity must be addressed to achieve the goal of data analytics. Challenges in analyzing data include developing methodologies, common task frameworks, and learning paradigms to handle data, domain, behavioral, social, and environmental complexity. Data scientists must be able to learn from heterogeneous sources and inputs, parallel and distributed inputs, and their infinite dynamics in real time; support on-the-fly active and adaptive learning of large data volumes in computational resource-poor environments (such as embedded sensors), as well as multisource learning, while considering the relations and interactions between sensors; enable combined learning across multiple learning objectives, sources, feature sets, analytical methods, frameworks, and outcomes; learn non-IID data-mixing coupling relationships with heterogeneity;[2] and ensure transparency and certainty of learning models and outcomes.

Other requirements for managing and exploiting data include appropriate design of experiments and mechanisms. Inappropriate learning could result in misleading or harmful outcomes, as in a classifier that works for balanced data but could mistakenly classify biased and sparse cases in anomaly detection.

The complexity of a deliverable data product, or "deliverable complexity" becomes an obstruction when actionable insight[7] is the focus of a data science application. Such complexity necessitates identification and evaluation of the outcomes that satisfy technical significance and have high business value from both an objective and a subjective perspective. The related challenges for data scientists also involve designing the appropriate evaluation, presentation, visualization, refinement, and prescription of learning outcomes and deliverables to satisfy diverse business needs, stakeholders, and decision support. In general, data deliverables to business users must be easy to understand and interpretable by nonprofessionals, revealing insights that directly inform and enable decision making and possibly having a transformative effect on business processes and problem solving.

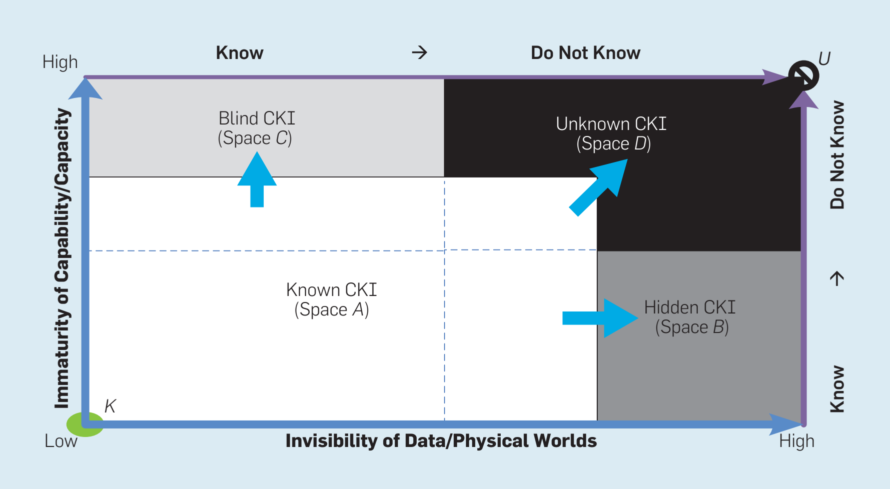
X-Intelligence in Data Science

Data science is a type of "intelligence science" that aims to transform data into knowledge, intelligence, and wisdom.[21] In this transformation, comprehensive intelligence,[3] or "X-intelligence," is often used to address a complex data science problem, referring to comprehensive and valuable information. X-intelligence can help inform the deeper, more structured and organized comprehension, representation, and problem solving in the underlying complexities and challenges.

Data intelligence highlights the most valuable information and narratives in the formation and solution of business problems or value in the corresponding data. Intelligence hidden in data is discovered by data science through its ability to understand data characteristics and complexities. Apart from the usual focus on complexities in data structures, distribution, quantity, speed, and quality, the focus in data science is on the intelligence hidden in the unknown "Space *D*" in Figure 2. For example, in addition to existing protocols for cancer treatment, determining what new and existing treatments fail on which patients might be informed by analyzing health-care data and diversified external data relevant to cancer patients. The level of data intelligence depends on how much and to what extent a data scientist is able to deeply understand and represent data characteristics and complexities.

Data scientists discover behavior intelligence by looking into the activities, processes, dynamics, and impact of individual and group actors, or the behavior and business quantifiers, owners, and users in the physical world. Such discovery requires they be able to bridge the gap between the data world and the physical world by connecting what happened and what will happen in the problem and discovering behavior insights through behavior informatics.[1] For example, in monitoring online shopping websites, regulators must be able to recognize whether ratings and comments are made by robots, rather than humans;

Figure 2. Known-to-unknown discovery in data science.



likewise, in social media, detecting algorithms, or robot-generated comments, in billions of daily transactions is itself a computational challenge. Constructing sequential behavior vector spaces and modeling interactions with other accounts in a given time period and then differentiating abnormal behaviors may be useful for understanding the difference between proactive and subjective human activity and the reactive and patternable behaviors of software robots.

Domain intelligence emerges from relevant domain factors, knowledge, meta-knowledge, and other domain-specific resources associated with a problem and its target data. Qualitative and quantitative domain intelligence can help inform and enable a data scientist's deep understanding of domain complexities and their roles in discovering unknown knowledge and actionable insight. For example, to learn high-frequency trading strategies for use with stock data, a strategy modeler must include the "order book" and microstructure of the related "limit market."

Human intelligence plays a central role in complex data science systems through explicit, or direct, involvement of human intuition, imagination, empirical knowledge, belief, intention, expectation, runtime supervision, evaluation, and expertise. It also concerns the implicit, or indirect, involvement of human intelligence in the form of imaginative thinking, emotional intelligence, inspiration, brainstorming, reasoning inputs, and embodied cognition, as in convergent thinking through interaction with fellow humans. For example, as "data-science thinking"[6] is crucial for addressing complex data problems, data scientists must be able to apply subjective factors, qualitative reasoning, and critical imagination.

Network intelligence emerges from both Web intelligence and broad-based networking and connected activities and resources, especially through the Internet of Things, social media, and mobile services. Information and facilities from the networks involved in target business problems can contribute useful information for complex data-science problem solving; a relevant example is crowdsourcing-based open source system development and algorithm design.

Organizational intelligence emerges from the proper understanding, involvement, and modeling of organizational goals, actors, and roles, as well as structures, behaviors, evolution and dynamics, governance, regulation, convention, process, and workflow in data science systems. For example, the cost effectiveness of enterprise analytics and functioning of data science teams rely on organizational intelligence.

Social intelligence emerges from the social complexities discussed earlier. Human social intelligence is embedded in social interactions, group goals and intentions, social cognition, emotional intelligence, consensus construction, and group decision making. Social intelligence is also associated with social-network intelligence and collective interactions among social systems, as well as the business rules, law, trust, and reputation for governing social intelligence. Typical artificial social systems include social networks and social media in which data-driven social complexities are understood through social-influence modeling, latent relation modeling, and community formation and evolution in online societies.

Environmental intelligence is also hidden in data science problems, as specified in terms of the underlying domain and related organizational, social, human, and network intelligence. Data science systems are open, with interactions between the world of transformed data and the physical world functioning as the overall data environment. Examples include context-aware analytics involving contextual factors and evolving interactions and changes between data and context, as in infinite-dynamic-relation modeling in social networks.

Known-to-Unknown Transformation

Complex data science problem-solving journeys taken by data scientists represent a cognitive progression from understanding known to unknown complexities in order to transform data into knowledge, intelligence, and insight for decision taking by inventing and applying respective data-intelligence discovery capabilities. In this context, knowledge represents processed information in terms of information mixture, procedural action, or propositional rules; resulting insight refers to the deep understanding of intrinsic complexities and mechanisms in data and its corresponding physical world.

Figure 2 outlines data science progression aiming to reduce the immaturity of capabilities and capacity (y-axis) to better understand the hidden complexities, knowledge, and intelligence (CKI) in data/physical worlds (x axis) from the 100% known state K to the 100% unknown state U. Based on data/physical world visibility and capability/capacity maturity, data science can be categorized into four data challenges:

"Space *A*" represents the known space; that is, "I (my mature capability/capacity) know what I know (about the visible world)." This is like the ability of sighted people to recognize an elephant by seeing the whole animal, whereas non-sighted people might be able to identify only part of the animal through touch. Knowledge concerning visible data is known to people with mature capability/capacity; that is, their capability/capacity maturity is sufficient to understand data/physical-world invisibility. This insight corresponds to well-understood areas in data science. Examples include profiling and descriptive analysis that applies existing models to data deemed by data analysts to follow certain assumptions.

"Space *B*" represents the hidden space; that is, "I know what I do not know (about the unseen world)." For some people or disciplines, even though certain aspects of their capability/capacity is mature, CKI is hidden to (so cannot be addressed by) current data science capability/capacity, thus requiring more-advanced capability/capacity. Examples include existing IID models (such as k-means and the k-nearest neighbors algorithm) that cannot handle non-IID data.

"Space *C*" represents the blind space; that is, "I (my immature capability) do not know what I know (about the world)." Although CKI is visible to some people or disciplines, their capability/capacity is also mature, but CKI and capability/capacity do not match well; immaturity thus renders them blind to the world. An example might be when even established social scientists try to address a data science problem.

"Space *D*" represents the unknown; that is, "I do not know what I do not know, so CKI in the hidden world is unknown due to immature capability." This is the area today on which data science focuses its future research and discovery. Along with increased invisibility, the lack of capability maturity also increases. In the world of fast-evolving big data, CKI invisibility increases, resulting in an ever-larger unknown space.

The current stage of data science capability and maturity, or "We do not know what we do not know," can be explained in terms of unknown perspectives and scenarios. As outlined in Figure 3, the unknown world presents "unknownness" in terms of certain definable categories, including problems and complexities; hierarchy, structures, distributions, relations, and heterogeneities; capabilities, opportunities, and gaps; and solutions.

Data Science Directions

Here, I consider the applied data science conceptual landscape, followed by two significant aspirational goals: non-IID data learning and human-like intelligence.

Data science landscape. The X-complexity and X-intelligence in complex data science systems and widening gap between world invisibility and capability/capacity immaturity yield new research challenges that motivate development of data science as a discipline. Figure 4 outlines the conceptual

Figure 3. Hidden world in data science.

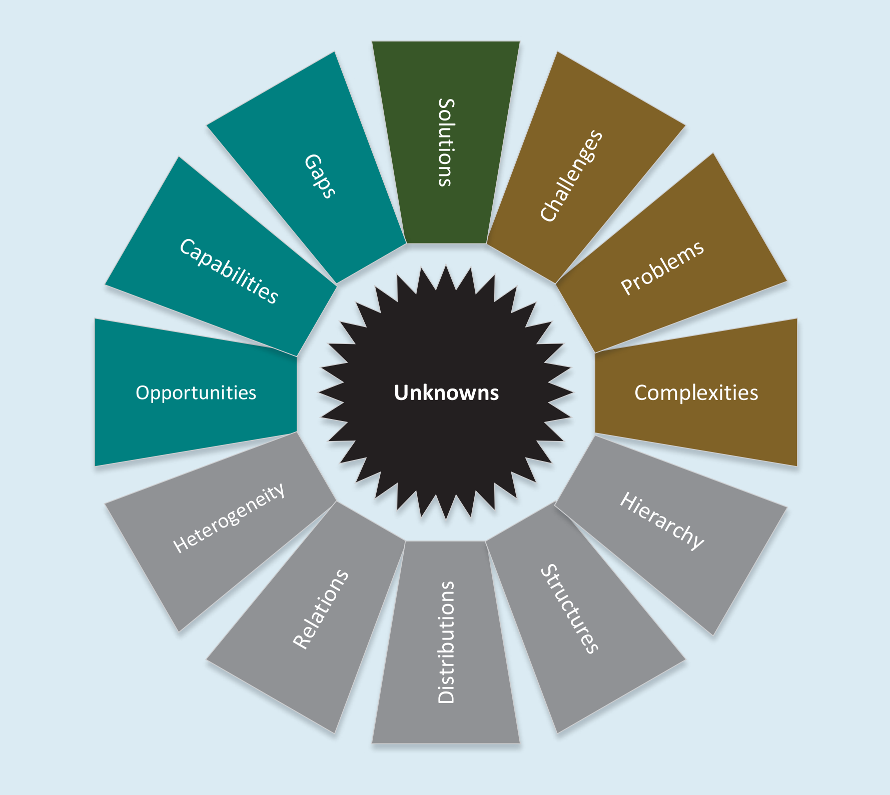
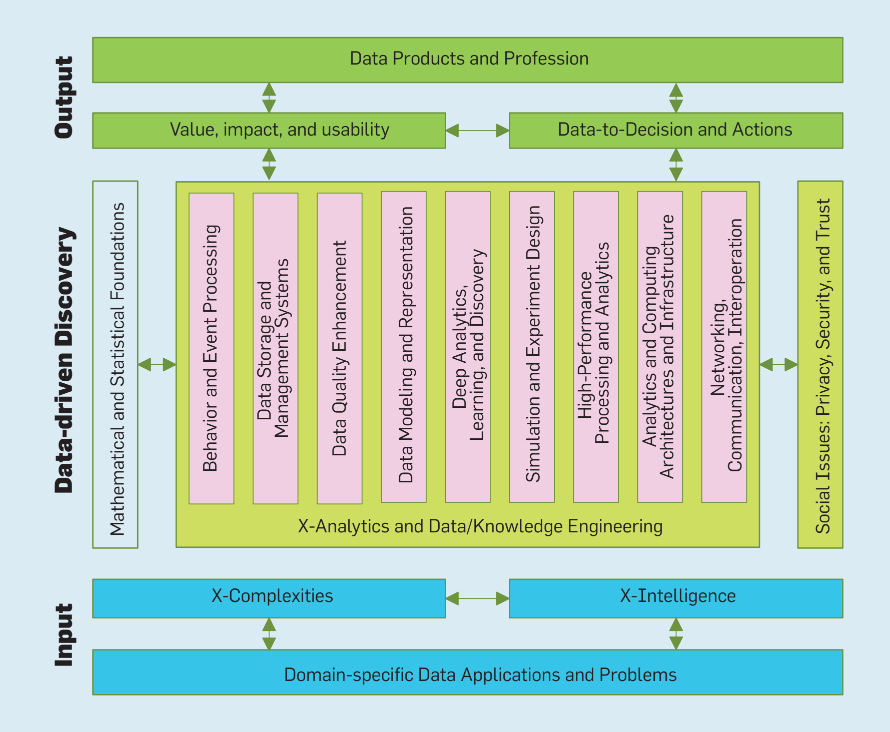


Figure 4. Data science conceptual landscape.


Output
Data Products and Profession
Value, impact, and usability
Data-to-Decision and Actions
Data-driven Discovery
Mathematical and Statistical Foundations
Behavior and Event Processing
Data Storage and Management Systems
Data Quality Enhancement
Data Modeling and Representation
Deep Analytics, Learning, and Discovery
Simulation and Experiment Design
High-Performance Processing and Analytics
Analytics and Computing Architectures and Infrastructure
Networking, Communication, Interoperation
X-Analytics and Data/Knowledge Engineering
Social Issues: Privacy, Security, and Trust
Input
X-Complexities
X-Intelligence
Domain-specific Data Applications and Problems

landscape of data science and its major research challenges by taking an interdisciplinary, complex-system-based, hierarchical view.

As in Figure 4, the data science landscape consists of three layers: "data input," including domain-specific data applications and systems, and X-complexity and X-intelligence in data and business problems; "data-driven discovery" consisting of discovery tasks and challenges; and "data output" consisting of results and outcomes.

Research challenges and opportunities emerge in all three in terms of five areas not otherwise managed well through non-data-science methodologies, theories, or systems:

Data/business understanding. The aim is for data scientists, as well as data users, to identify, specify, represent, and quantify the X-complexities and X-intelligence that cannot be managed well through existing theories and techniques but nevertheless are embedded in domain-specific data and business problems. Examples include how to understand in what forms, at what level, and to what extent the respective complexities and intelligence interact with one another and to devise methodologies and technologies for incorporating them into data science tasks and processes;

Mathematical and statistical foundation. The aim is to enable data scientists to disclose, describe, represent, and capture complexities and intelligence for deriving actionable insight. Existing analytical and computational theories may need to be explored as to whether, how, and why they are insufficient, missing, or problematic, then extended or redeveloped to address the complexities in data and business problems by, say, supporting multiple, heterogeneous, large-scale hypothesis testing and survey design, learning inconsistency, and uncertainty across multiple sources of dynamic data. Results might include deep representation of data complexities, large-scale, fine-grain personalized predictions, support for non-IID data learning, and creation of scalable, transparent, flexible, interpretable, personalized, parameter-free modeling;

Data/knowledge engineering and X-analytics. The aim is to develop domain-specific analytic theories, tools, and systems not available in the relevant body of knowledge to represent, discover, implement, and manage the data, knowledge, and intelligence and support the corresponding data and analytics engineering. Examples include automated analytical software that automates selection and construction of features and models, as well as the analytics process in its self-understanding of intrinsic data complexities and intelligence, and that self-monitors, self-diagnoses, and self-adapts to data characteristics, domain-specific context and learning objectives and potential, and learns algorithms that recognize data complexities and self-trains the corresponding optimal models customized for the data and objectives;

The metasynthesis of X-complexities and X-intelligence in complex data science problems might ultimately produce even super machine intelligence.

Data quality and social issues. The aim here is to identify, specify, and respect social issues in domain-specific data, business-understanding, and data science processes, including use, privacy, security, and trust, and make possible social issues-based data science tasks not previously handled well. Examples include privacy-preserving analytical algorithms and benchmarking the trustworthiness of analytical outcomes;

Data value, impact, utility. The aim is to identify, specify, quantify, and evaluate the value, impact, and utility of domain-specific data that cannot be addressed through existing measurement theories and systems. Examples involve data actionability, utility, and value; and

Data-to-decision and action-taking challenges. The aim is to develop decision-support theories and systems to enable data-driven decisions and insight-to-decision transformation, incorporating prescriptive actions and strategies that cannot be managed through existing technologies and systems. Examples include ways to transform analytical findings into decision-making strategies.

Since data/knowledge engineering and advanced analytics[6] play a key role in data science, the focus is on specific research questions not previously addressed. Data-quality enhancement is fundamental to handling data-quality issues like noise, uncertainty, missing values, and imbalance that may be present due to the increasing scale of complexity and data-quality issues (such as cross-organizational, cross-media, cross-cultural, and cross-economic mechanisms) emerging in the big-data and Internet-based data/business environment.

Data scientists seek to model, learn, analyze, and mine data, including X-complexities and X-intelligence. For example, being able to perform deep analytics is essential for discovering unknown knowledge and intelligence in the unknown space in Figure 2 that cannot be handled through existing latent learning and descriptive and predictive analytics; another option might be to integrate data-driven and model-based problem solving, balancing common learning models and frameworks and domain-specific data complexities and intelligence-driven evidence learning.

X-complexity and X-intelligence pose additional challenges to simulation and experimental design, including how to simulate the complexities, intelligence, working mechanisms, processes, and dynamics in data and corresponding business systems and how to design experiments to explore the effect of business managers' data-driven decisions. Big-data analytics requires high-performance processing and analytics that support large-scale, real-time, online, high-frequency, Internet-based, cross-organizational data processing and analytics while balancing local and global resource objectives. Such an effort may require new distributed, parallel, high-performance infrastructure, batch, array, memory, disk, and cloud-based processing and storage, data-structure-and-management systems, and data-to-knowledge management.

Complex data science also challenges existing analytics and computing architectures and infrastructure to, say, invent analytics and computing architectures and infrastructure based on memory, disk, cloud, and Internet resources. Another important issue for developers of data systems is how to support the networking, communication, and interoperation of the various data science roles within a distributed data science team. Such coordination requires distributed cooperative management of projects, data, goals, tasks, models, outcomes, work flows, task scheduling, version control, reporting, and governance.

Addressing them involves systematic and interdisciplinary approaches possibly requiring synergy among many related research areas. Such synergy is due to taking on complex data science problems that cannot be addressed through one-off efforts. For instance, data structures, computational infrastructure, and detection algorithms are required for high-frequency real-time risk analytics in extremely large online businesses like electronic commerce and financial trading.

Violating assumptions in data science. Big data includes X-complexities, including complex coupling relationships and/or mixed distributions, formats, types and variables, and unstructured and weakly structured data. Complex data poses significant challenges to many mathematical, statistical, and analytical methods built on relatively narrow assumptions, owing to the fact that they are routinely violated in big-data analytics. When assumptions are violated, modeling outcomes may be inaccurate, distorted, or misleading. In addition to general scenarios (such as whether data violates the assumptions of normal distribution, t-test, and linear regression), an assumption check applies to broad aspects of a business problem's data, including independence, normality, linearity, variance, randomization, and measurement that apply to population data and analysis.

Fundamental work on detecting and verifying such validations is limited, and even less has sought to invent new theories and tools to manage and circumvent assumption violations in big data. One such violation I highlight here is the IID assumption, as big, complex data (referring to objects, attributes, and values[2]) is essentially non-IID, whereas most existing analytical methods are IID.[2]

In a non-IID data problem (see Figure 5a), non-IIDness, as outlined in Figure 5c, refers to the mixture of couplings, including co-occurrence, neighborhood, dependence, linkage, correlation, and causality, and other poorly explored and unquantified relations involving, say, sophisticated cultural and religious connections and influence, as well as heterogeneity within and between two or more aspects of a data system (such as entity, entity class, entity property like a variable, process, fact, and state of affairs) or other types of entities or properties (such as learning algorithms, and learned results) appearing or produced prior to, during, or after a target process (such as a learning task). By contrast, IIDness essentially ignores or simplifies all these properties, as outlined in Figure 5b.

Learning visible and especially invisible non-IIDness is fundamental for data scientists looking for deep understanding of data with weak and/or un-

Figure 5. IIDness vs. non-IIDness in data science problems.

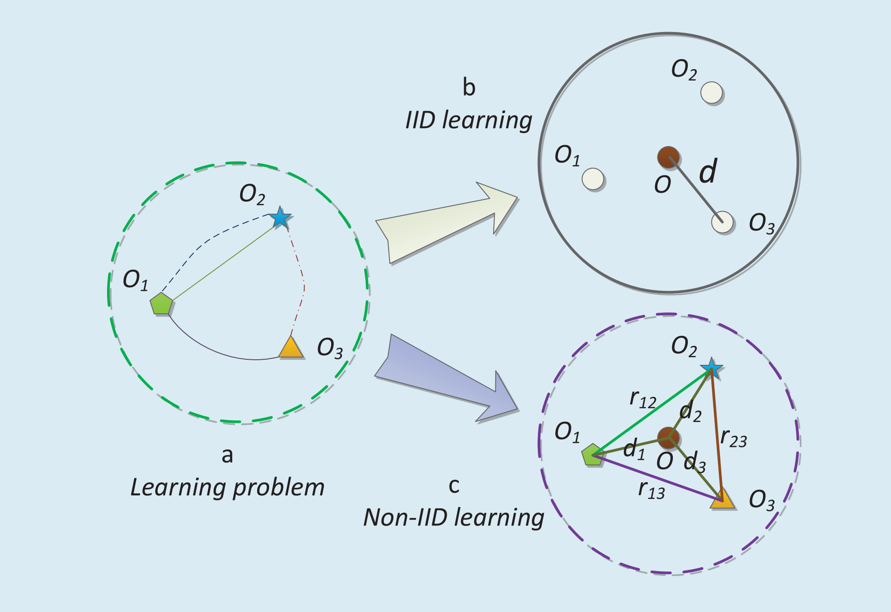


Figure 6. Synthesizing X-intelligence in data science.

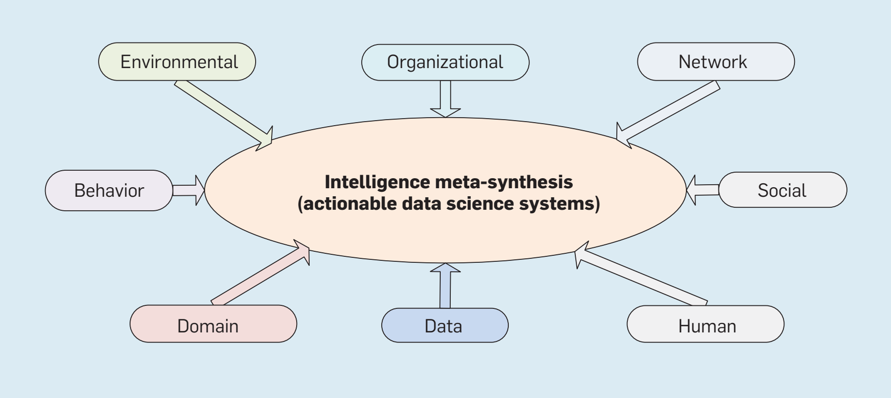
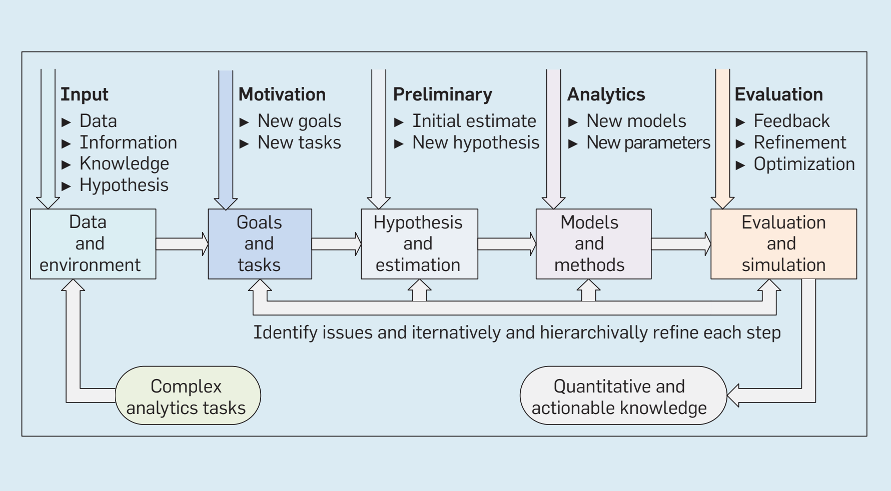


Figure 7. Complex data science problems: qualitative-to-quantitative X-intelligence metasynthesis.



clear structures, distributions, relationships, and semantics. In many cases, locally visible but globally invisible (or vice versa) non-IIDness takes a range of forms, structures, and layers on diverse entities. Individual learners cannot tell the whole story due to their inability to identify such complex non-IIDness. Effectively learning the widespread, visible, and invisible non-IIDness of big data is crucial for data scientists trying to gain a complete picture of an underlying business problem.

Data analysts often focus on learning explicit non-IIDness, or visible and easy to learn. The hybridization of multiple analytical methods on combinations of multiple sources of data into a big table for analysis typically falls into this category of non-IID systems. Computing non-IIDness refers to understanding, formalizing, and quantifying the non-IID aspects of data[2] (such as entities, interactions, layers, forms, and strength of non-IIDness). Non-IID learning systems seek to understand non-IIDness in data-analytics systems, from values, attributes, objects, methods, and measures to processing outcomes (such as mined patterns).

I now explore the prospects for inventing new data science theories and tools for non-IIDness and non-IID data learning,[2] including how to address non-IID data characteristics (not just variables), in terms of new feature analysis:

Deep understanding of non-IID data characteristics. The aim is to identify, specify, and quantify non-IID data characteristics, factors, types, and levels of non-IIDness in data and business, and identify the difference between what can be captured and what cannot be captured through existing technologies;

Non-IID feature analysis and construction. The aim is to invent new theories and tools for analyzing feature relationships by considering non-IIDness within and between features and objects and developing new theories and algorithms for selecting, mining, and constructing features;

Non-IID learning theories, algorithms, and models. The aim is to create new theories, algorithms, and models for analyzing, learning, and mining non-IID data by considering various couplings and heterogeneity; and

Non-IID similarity and evaluation metrics. The aim is to develop new similarity and dissimilarity learning methods and metrics, as well as evaluation metrics that consider non-IIDness in data and business.

More broadly, many existing data-oriented theories, designs, mechanisms, systems, and tools may have to be reinvented in light of non-IIDness. In addition to incorporating non-IIDness into data mining, machine learning, and general data analytics, non-IIDness is found in other well-established bodies of knowledge, including mathematical and statistical foundations, descriptive-analytics theories and tools, data-management theories and systems, information-retrieval theories and tools, multimedia analysis, and various X-analytics.[6]

Data characteristics and X-complexities. To address critical issues in data-driven discovery like assumption violations, I assume data characteristics and X-complexities determine the values, complexities, and quality of data-driven discovery. Data characteristics refer to the profile and complexities of data (generally a dataset) that can be described in terms of data factors (such as distribution, structure, hierarchy, dimension, granularity, heterogeneity, and uncertainty).

Understanding data characteristics and X-complexities involves four fundamental data science challenges and directions:[6] definition of data characteristics and X-complexities; how to represent and model data characteristics and X-complexities; data-characteristics- and X-complexities-driven data understanding, analysis, learning, and management; and how to evaluate the quality of data understanding, analysis, learning, and management in terms of data characteristics and X-complexities. Unfortunately, only limited theories and tools are available for addressing them.

Data-brain and human-like ma-

chine intelligence. Computer scientists, economists, and politicians, as well as the general public, debate whether and when machines might replace humans.[22] While it may not be possible to build data-brain or thinking machines with human-like abilities, data science, especially big-data analytics, is driving a technological revolution, from implementing logical-thinking-centered machine intelligence to creative-thinking-oriented machine intelligence. It may be partially reflected in Google's AlphaGo (https://deepmind.com/) defeat of top-ranked Chinese Go player Ke Jie in 2017 and South Korean grandmaster Lee Sedol in 2016, as well as the Facebook emotion experiment,[14] but none has actually exhibited human-like imagination or thinking. This revolution (such as through data science thinking[6]), if truly able to mimic human intelligence, may transform machine intelligence, changing the human-machine separation of responsibilities.

Curiosity is a critical human capability, starting the moment we are born. We want to know what, how, and why everything. Curiosity connects other cognitive activities, particularly imagination, reasoning, aggregation, creativity, and enthusiasm to produce new ideas, observations, concepts, knowledge, and decisions. Humans manage to upgrade their own intelligence through experience, exploration, learning, and reflection. Accordingly, a critical goal for data scientists is to enable data- and X-intelligence-driven machines to generate, retain, and simulate human curiosity through learning inquisitively from data and X-intelligence.

Imaginative thinking differentiates humans from machines designed with sense-effect, learning, reasoning, and optimization mechanisms. Human imagination is intuitive, creative, evolving, and uncertain. It also represents a great yet challenging opportunity for transforming logic, patterns, and predefined sense-effect-mechanisms-driven machines into human-like data systems. Such a transformation would require machines able to simulate human-imagination processes and mechanisms. Existing knowledge representation, aggregation, computational logic, reasoning, and logic thinking incorporated into machines may never quite deliver machine curiosity, intuition, or imagination. Existing data and computer theories, operating systems, system architectures and infrastructures, computing languages, and data management must still be fundamentally reformed by, say, simulating, learning, reasoning, and synthesizing original thoughts from cognitive science, social science, data science, and intelligence science to render machines creative. They also must be able to engage X-intelligence in a non-predefined, "non-patternable" way, unlike existing simulation, learning, and computation, which are largely predefined or design-based by default.

To enable discovery, data-analytical thinking, a core aspect of data science thinking,[6] needs to be built into data products and learned by data professionals. Data-analytical thinking is not only explicit, descriptive, and predictive but also implicit and prescriptive. Complex data problem solving requires systematic, evolving, imaginative, critical, and actionable data science thinking. In addition to computational thinking, a machine might ultimately be able to mimic human approaches to information processing by synthesizing comprehensive data, information, knowledge, and intelligence through cognitive-processing methods and processes.

Developing Complex Systems

The X-complexities and X-intelligence discussed earlier render a complex data system equivalent to an open complex intelligent system.[3] Use of X-intelligence by a data scientist could take one of two paths: "single intelligence engagement" or "multi-aspect intelligence engagement." An example of the former is domain knowledge in data analytics and user preferences in recommender systems. Single-intelligence engagement applies to simple data science problem solving and systems. In general, multi-aspect X-intelligence can be found in complex data science problems.

As outlined in Figure 6, the performance of a data science-problem-solving system depends on recognition, acquisition, representation, and integration of relevant X-intelligence. To enable an X-intelligence-driven complex data science problem-solving process, data scientists need new methodologies and system-engineering methods. The theory of "metasynthetic engineering"[3,20] and integration of ubiquitous intelligence might provide useful guidance for synthesizing X-intelligence in complex data and business systems.

The principle of "intelligence metasynthesis" of multiple types of intelligence[3,20] involves, synthesizes, and uses ubiquitous intelligence in the complex data environment to understand the nature of data and related problems, invent discovery systems, discover interesting knowledge, and generate actionable insights.[7] Intelligence metasynthesis applies to solving complex data science problems involving complex system engineering in which multiple aspects of complexity and intelligence may be embedded in the data, environment, and problem-solving process. The "reductionism" methodology[3] for data and knowledge exploration may not work well because the problem may not be clear, specific, and quantitative so cannot be decomposed and analyzed effectively. In contrast, analysis through a holistic lens does not equal the sum of the analysis of the parts, a common challenge developing complex systems.[20]

Accordingly, in light of the theory of "system complexities" and corresponding methodologies "systematism,"[3,20] a methodology that synthesizes reductionism and "holism"[6,7] may then be more applicable for analyzing, designing, and evaluating complex data problems.

When a data science problem involves large-scale data objects, multiple levels of subtasks, objects, sources, and types of data from online, business, mobile, or social networks, complicated contexts, human involvement and domain constraints, the problem thus reflects the characteristics of an open complex system.[3,20] The problem is also likely to involve common system complexities, including openness, scale, hierarchy, human involvement, societal characteristics, dynamic characteristics, uncertainty, and imprecision.[3,19,20]

Although specific big-data analytical tasks are manageable by following existing analytical methodologies, typical cross-enterprise, global, and Internet-based data science projects (such as global financial crisis and terrorist activities) satisfy most if not all such complexities. This level of complex data science involves X-complexities problems, and their resolution must first synthesize the X-intelligence in the problems. One approach to instantiate the systematism methodology is “qualitative-to-quantitative metasynthesis,”[3,20] as proposed by Chinese scientist Xuesen Qian (also known as Hsue-Shen Tsien) to guide system engineering in large-scale open systems.[20] Such qualitative-to-quantitative metasynthesis supports exploration of open complex systems through an iterative cognitive and problem-solving process on a human-centered, human-machine-cooperative problem-solving platform in which human, data, and machine intelligence, along with X-intelligence, must be engaged, quantified, and synthesized. Implementing it for open complex intelligent systems, the “metasynthetic computing and engineering” (MCE) approach[3] provides a systematic computing and engineering guide and suite of system-analysis tools.

Figure 7 outlines the process of applying the qualitative-to-quantitative metasynthesis methodology to complex data science problems. MCE supports an iterative, hierarchical problem-solving process, incorporating internal and external inputs, including data, information, domain knowledge, initial hypotheses, and underlying environmental factors. Data scientists would start by presetting analytics goals and tasks to be explored on the given data by incorporating domain, organizational, social and environmental complexities and intelligence. They would then use preliminary observations obtained from domain and experience to identify and verify qualitative and quantitative hypotheses and estimations that guide development of modeling and analytics methods. Findings would then be evaluated and fed back to the corresponding procedures for refining and optimizing understanding of previously unknown problem challenges, goals, and discovery methods. Following these iterative and hierarchical steps toward qualitative-to-quantitative intelligence transformation would thus disclose and quantify the initial problem “unknownness.” Finally, actionable knowledge and insight would be identified and delivered to businesspeople who would address data complexities and business goals.

As an example of how to deliver actionable knowledge, domain-driven data mining[7] aims to integrate X-intelligence and X-complexities for complex knowledge-discovery problems. Domain-driven data mining advocates a comprehensive process of synthesizing data intelligence with other types of intelligence to prompt new intelligence to address gaps in existing data-driven methods, delivering actionable knowledge to business users. The metasynthesis of X-complexities and X-intelligence in complex data science problems might ultimately produce even super machine intelligence. Super-intelligent machines could then understand, represent, and learn X-complexities, particularly data characteristics; acquire and represent unstructured, ill-structured, and uncertain human knowledge; support involvement of business experts in the analytics process; acquire and represent imaginative and creative thinking in group heuristic discussions among human experts; acquire and represent group/collective interaction behaviors; and build infrastructure involving X-intelligence. While a data brain cannot mimic special human imagination, curiosity, and intuition, the simulation and modeling of human behavior and human-data systems interaction and cooperation promise to approach human-like machine intelligence.

Conclusion

The low-level X-complexities and X-intelligence characterizing complex data science problems reflect the gaps between the world of hidden data and existing data science immaturity. Filling them requires a disciplinewide effort to build complex data science thinking and corresponding methodologies from a complex-system perspective. The emerging data science evolution means opportunities for breakthrough research, technological innovation, and a new data economy. If parallels are drawn between evolution of the Internet and evolution of data science, the future and the socioeconomic and cultural impact of data science will be unprecedented indeed, though as yet unquantifiable.

References

1. Cao, L.B. In-depth behavior understanding and use: The behavior informatics approach. *Information Science 180*, 17 (Sept. 2010), 3067–3085.
2. Cao, L.B. Non-IIDness learning in behavioral and social data. *The Computer Journal 57*, 9 (Sept. 2014), 1358–1370.
3. Cao, L.B. *Metasynthetic Computing and Engineering of Complex Systems*. Springer-Verlag, London, U.K., 2015.
4. Cao, L.B. Data science: Nature and pitfalls. *IEEE Intelligent Systems 31*, 5 (Sept.-Oct. 2016), 66–75.
5. Cao, L.B. Data science: A comprehensive overview. *ACM Computing Surveys* (to appear).
6. Cao, L.B. *Understanding Data Science*. Springer, New York (to appear).
7. Cao, L.B., Yu, P.S., Zhang, C., and Zhao, Y. *Domain Driven Data Mining*. Springer, Springer-Verlag, New York, 2010.
8. Cao, L.B., Yu, P.S., and Kumar, V. Nonoccurring behavior analytics: A new area. *IEEE Intelligent Systems 30*, 6 (Nov. 2015), 4–11.
9. Cleveland, W.S. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review 69*, 1 (Dec. 2001), 21–26.
10. Diggle, P.J. Statistics: A data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*, 4 (Oct. 2015), 793–813.
11. Donoho, D. *50 Years of Data Science*. Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, 2015; http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
12. Huber, P.J. *Data Analysis: What Can Be Learned from the Past 50 Years*. John Wiley & Sons, Inc., New York, 2011.
13. Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., and Shahabi, C. Big data and its technical challenges. *Commun. ACM 57*, 7 (July 2014), 86–94.
14. Kramer, A.D., Guillory, J.E., and Hancock, J.T. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences 111*, 24 (Mar. 2014), 8788–8790.
15. Lazer, D., Kennedy, R., King, G., and Vespignani, A. The parable of Google flu: Traps in big data analysis. *Science 343*, 6176 (Mar. 2014), 1203–1205.
16. Manyika, J. and Chui, M. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute, 2011.
17. Matsudaira, K. The science of managing data science. *Commun. ACM 58*, 6 (June 2015), 44–47.
18. Mattmann, C.A. Computing: A vision for data science. *Nature 493*, 7433 (Jan. 24, 2013), 473–475.
19. Mitchell, M. *Complexity: A Guided Tour*. Oxford University Press, Oxford, U.K., 2011.
20. Qian, X., Yu, J., and Dai, R. A new discipline of science—The study of open complex giant system and its methodology. *Journal of Systems Engineering and Electronics 4*, 2 (June 1993), 2–12.
21. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information and Communication Science 33*, 2 (Apr. 2007), 163–180.
22. Suchma, L. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press, Cambridge, U.K., 2006.
23. Tukey, J.W. The future of data analysis. *The Annals of Mathematical Statistics 33*, 1 (Mar. 1962), 1–67.
24. Tukey, J.W. *Exploratory Data Analysis*. Pearson, 1977.

Longbing Cao (longbing.cao@gmail.com) is a professor in the Advanced Analytics Institute at the University of Technology Sydney, Australia.