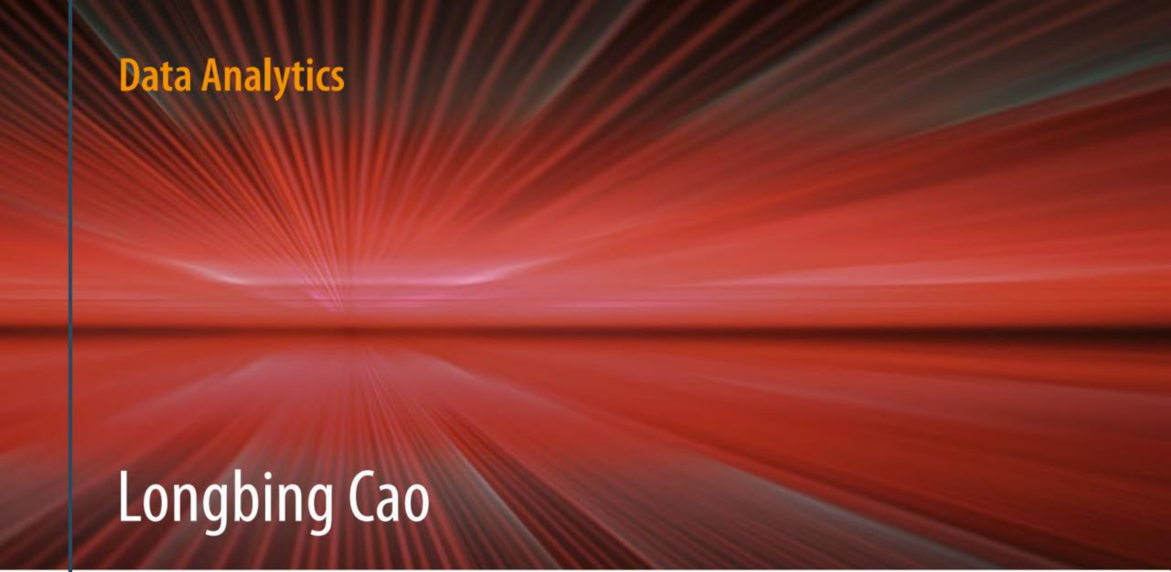
Data Analytics



Longbing Cao

Data Science Thinking

The Next Scientific, Technological and Economic Revolution

Springer

Longbing Cao

Data Science Thinking

The Next Scientific, Technological and Economic Revolution

Springer

Longbing Cao
Advanced Analytics Institute
University of Technology Sydney
Sydney, NSW, Australia

ISSN 2520-1859 ISSN 2520-1867 (electronic)
Data Analytics
ISBN 978-3-319-95091-4 ISBN 978-3-319-95092-1 (eBook)
https://doi.org/10.1007/978-3-319-95092-1

Library of Congress Control Number: 2018952348

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Concepts and Thinking

Part III Industrialization and Opportunities

Part I
Concepts and Thinking

Chapter 1
The Data Science Era

1.1 Introduction

We are living in the age of big data, advanced analytics, and data science. The trend of "big data growth" [29, 106, 266, 288, 413] (data deluge [210]) has not only triggered tremendous hype and buzz, but more importantly presents enormous challenges, which in turn have brought incredible innovation and economic opportunities.

Big data has attracted intense and growing attention from major governmental organizations, including the United Nations [399], USA [407], EU [101] and China [196], traditional data-oriented scientific and engineering fields, as well as non-traditional data engineering domains such as social science, business and management [91, 252, 265, 472].

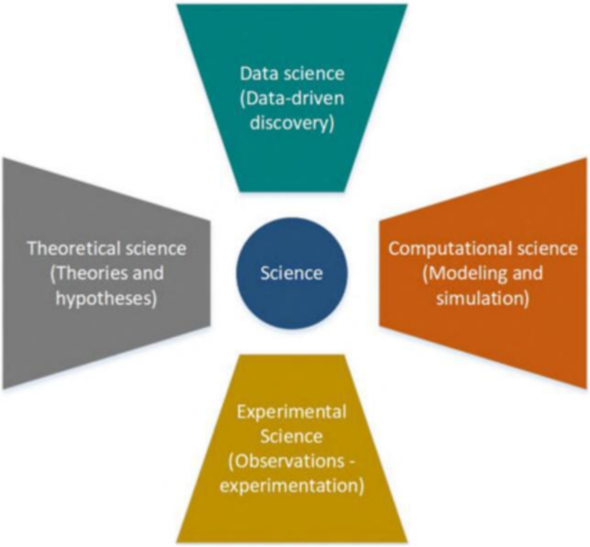
From the disciplinary development perspective, recognition of the significant challenges, opportunities and values of big data is fundamentally reshaping traditional data-oriented scientific and engineering fields. It is also reshaping non-traditional data engineering domains such as social science, business and management [91, 252, 265, 472]. This paradigm shift is driven not just by data itself but by the many other aspects of the power of data (simply *data power*), from data-driven science to data-driven economy, that could be created, invented, transformed and/or adjusted by understanding, exploring and utilizing data.

This trend and its potential have triggered new debate about *data-intensive scientific discovery* as a new paradigm, the so-called "fourth science paradigm", which unifies experiment, theory and computation (corresponding to *empirical science* or *experimental science*, *theoretical science* and *computational science*) [198, 209], as shown in Fig. 1.1. Data is regarded as the new Intel Inside [319], or the new oil and strategic asset, and is driving—even determining—the future of science, technology, the economy, and possibly everything else in our world.

In 2005 in Sydney, we were asked a critical question at a brainstorming meeting about data science and data analytics by several local industry representatives from

L. Cao, *Data Science Thinking*, Data Analytics,
https://doi.org/10.1007/978-3-319-95092-1_1

Fig. 1.1 Four scientific paradigms



major analytics software vendors: "Information science has been there for so long, why do we need data science?" Related fundamental questions often discussed in the community include "What is data science?" [279], and "Is data science old wine in new bottles?" [2]. Data science and associated topics have become the key concern in panel discussions at conferences in statistics, data mining, and machine learning, and more recently in big data, advanced analytics, and data science. Typical topics such as "grand challenges in data science", "data-driven discovery", and "data-driven science" have frequently been visited and continue to attract wide and increasing attention and debate. These questions are mainly posited from research and disciplinary development perspectives, but there are many other important questions, such as those relating to data economy and competency, that are less well considered in the conferences referred to above.

A fundamental trigger for these questions and many others not mentioned here is the exploration of new or more complex challenges and opportunities [54, 64, 233, 252] in data science and engineering. Such challenges and opportunities apply to existing fields, including statistics and mathematics, artificial intelligence, and other relevant disciplines and domains. They are issues that have never been adequately addressed, if at all, in classic methodologies, theories, systems, tools, applications and economy. Such challenges and opportunities cannot be effectively accommodated by the existing body of knowledge and capability set without the development of a new discipline.

On the other hand, data science is at a very early stage and, apart from engendering enormous hype, it also causes a level of bewilderment, since the issues and possibilities that are unique to data science and big data analytics are not clear, specific or certain. Different views, observations, and explanations—some of them controversial—have thus emerged from a wide range of perspectives.

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

This chapter therefore presents an overview of the *data science era*, which incorporates the following aspects:

- Features of the data science era;
- The data science journey from data analysis to data science;
- The main driving forces of data-centric thinking, innovation and practice;
- The interest trends demonstrated in Internet search;
- Major initiatives launched by governments; and
- Major initiatives on the scientific agenda launched by scientific organizations.

The goal of this chapter is to present a comprehensive high level overview of what has been going on in communities that are representative of the data science era, before addressing more specific aspects of data science and associated perspectives in the remainder of the book.

1.2 Features of the Data Era

1.2.1 Some Key Terms in Data Science

Before proceeding to discuss the many aspects of data science, we list several key terms that have been widely accepted and discussed in relevant communities in relation to the data science era: data analysis, data analytics, advanced analytics, big data, data science, deep analytics, descriptive analytics, predictive analytics, and prescriptive analytics. These terms are highly connected and easily confused, and they are also the key terms widely used in the book. Table 1.1 thus lists and explains these terms.

A list of data science terminology is available at www.datasciences.info.

1.2.2 Observations of the Data Era Debate

With their emergence as significant new areas and disciplines, big data [25, 288] and data science [388] have been the subject of increased debate and controversy in recent years.

Table 1.1 Key terms in data science

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

After reviewing [63] a large number of relevant works in the literature that directly incorporate data science in their titles, we make the following observations about the big data buzz and data science debate:

- Very comprehensive discussion has taken place, not only within data-related or data-focused disciplines and domains, such as statistics, computing and informatics, but also in non-traditional data-related fields and areas such as social science and management. Data science has clearly emerged as an inter-, cross- and trans-disciplinary new field.
- In addition to the thriving growth in academic interest, industry and government organizations have increasingly realized the value and opportunity of data-driven innovation and economy, and have thus devised policies and initiatives to promote data-driven intelligent systems and economy.

- Although many discussions and publications are available, most (probably more than 95%) essentially concern existing concepts and topics discussed in statistics, artificial intelligence, pattern recognition, data mining, machine learning, business analytics and broad data analytics. This demonstrates how data science has developed and been transformed from existing core disciplines, in particular, statistics, computing and informatics.
- While data science as a term has been increasingly used in the titles of publications, it seems that a great many authors have done this to make the work look ‘sexier’. The abuse, misuse and over-use of the term “data science” is ubiquitous, and essentially contribute to the buzz and hype. Myths and pitfalls are everywhere at this early, and somehow impetuous, stage of data science.
- Very few thoughtful articles are available that address the low-level, fundamental and intrinsic complexities and problematic nature of data science, or contribute deep insights about the intrinsic challenges, directions and opportunities of data science as a new field.

It is clear that we are living in the era of big data and data science—an era that exhibits iconic features and trends that are unprecedented and epoch-making.

1.2.3 Iconic Features and Trends of the Data Era

In the era of data science, an essential question to ask is *what typifies this new era?* It is critical to identify the features and characteristics of the data science era. However, it is very challenging to provide a precise summary at this early stage.

To give a fair summary, the main characteristics of the data science era are discussed from the perspective of the transformation and paradigm shift caused by data science, the core driving forces, and the status of several typical issues confronting the data science field.

A data-centric perspective is taken to summarize the main characteristics of data science-related government initiatives, disciplinary development, economy, and profession, as well as the relevant activities in these fields, and the progress made to date.

We summarize eight data era features in Table 1.2 which represent this new age of science, profession, economy and education.

Data existence—Datafication is ubiquitous, and data quantification is ever-increasing: Data is physically, increasingly and ubiquitously generated at any time by any means. This goes beyond the traditional main sources of datafication [19]: sensors and management information systems. Today’s datafication devices and systems are everywhere, involved in and related to our work, study, entertainment, socio-cultural environment, and quantified personal devices and services [96, 143, 160, 363, 377, 462]. In addition, *data quantification is ever-increasing*: The data deluge features an exponential increase in the volume and variety of data at a speed

Table 1.2 Key features and trends of the data science era

Landmark	Significance
Physical existence	Datafication is ubiquitous, and data quantification is ever-increasing
Complexities	Data complexities cannot be handled by classic theories and systems
Strategic values	Data becomes a strategic asset Openness becomes a new paradigm and fashion
Research and development	Data science research and innovation drive a new scientific agenda
Startup business	Data-driven strategic data initiatives and startups start to dominate new business
Job market	Data scientist becomes a new profession
Business and economy	Data drives both the new data economy and traditional industry transformation
Disciplinary maturity	Data science becomes a new discipline, and data science is interdisciplinary

and in forms that were not previously imaginable and cannot be precisely predicted. There is apparently no end to this ever-increasing data quantification trend.

Data complexities—Data complexities cannot be handled by classic theories and systems: Data that is substantially complex cannot be well addressed by existing statistical and mathematical theories and systems, information theories and tools, analytics and learning systems.

Data value—Data becomes a strategic asset, and openness becomes a new paradigm and fashion: The strategic value of data is increasingly recognized by data owners and data generators, who treat data as a strategic asset for commercialization and other purposes. At the same time, there is a strong push for data to be open. Open source software, services and applications, and free repositories and services are a highlight of this data era. To a certain extent, open data and the open source environment are key drivers of the big data and data science era, propelling the spread of open data, open access, and open source to open innovation and open science, creating a new paradigm for research and science.

Data research and development—Data science research and innovation drives a new scientific field: Due to the significant data complexities and data values that have not been addressed in existing scientific and innovation systems, data science research and innovation is high on the current scientific agenda. More and more national science foundations, science councils, research foundations, and research and innovation policy-makers are increasing their funding support for data science innovation and basic research in both general scientific disciplines and specific areas such as health informatics, bioinformatics, and brain informatics.

Data startup—Data-driven strategic data initiatives and startups start to dominate the new business: We are seeing rapidly increasing strategic initiatives established by increasing numbers of governments, vendors, professional bodies, and large and small businesses. Data industrialization is driving the new wave of economic transformation and startups.

Data science job positioning—Data scientist becomes a new profession: Data science jobs dominate the job market, demonstrating a rapidly increasing trend which is marked by a high average salary. New data professional communities are formed, as evidenced by the creation of new chief officer roles such as chief data officer, chief analytics officer, and chief data scientist, as well as multiple roles which are broadly termed 'data scientist'. This leads to a business-driven, fast-growing, open data science community, and the development of various analytics customized for specific domains, such as agricultural analytics and social analytics.

Data economy—Data is driving both the new data economy and traditional industry transformation: This is not only represented by the emergence of data-focused companies and startups such as Google, Facebook, and Cloudera, but also by the data-driven transformation of traditional industry and core business, in particular, banking, capital markets, telecommunication, manufacturing, the food industry, healthcare business, medicine and medical services, and the educational sector. In addition to the above typical data-driven businesses, data industrialization is changing the Internet landscape, driving new data products, data systems, and data services that are embedded in social media, mobile applications, online business, and the Internet of Things (IoT). In core business and traditional industry, the changes result from data-based competition, productivity elevation, service enhancement, and decision efficiency and effectiveness, which, while not as visible as the new data economy, are just incredible and hitherto unimaginable.

Data science discipline—Data science becomes a new discipline, and data science is interdisciplinary: Universities, research institutions, vendors and commercial companies have rapidly recognized data science as a new discipline and are establishing an enormous number of awarded degrees, training courses, and online courses which are combined with existing interdisciplinary subjects from undergraduate level to doctoral level, or from non-award training programs. The last 5 years have seen a rapid increase in the creation of institutes, centers, and departments focusing on data science research, teaching and engagement across a broad range of international communities and research, government and industry agendas.

1.3 The Data Science Journey

This section summarizes the findings of a comprehensive survey in [63] and other related work, such as in [129, 172, 330], of the data science journey from data analysis to data science and the evolution of the interest in data science.

When was "data science" as a term first introduced? It is likely that the first appearance of "data science" as a term in literature was in the preface to Naur's book "Concise Survey of Computer Methods" [301] in 1974. In that preface, *data science* was defined as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences." Another term, "datalogy", had previously been introduced in 1968 as "the science of data and of data processes" [300]. These definitions are clearly more specific than those we discuss today. However, they have inspired today's significant move toward the comprehensive exploration of scientific content and development.

The past 50+ years have seen the transformation from data analysis to data science, and the trend is becoming more evident, widespread and profound. This evolutionary journey from data analysis [216] to data science started in the statistics and mathematics community in 1962. At that time, it was stated that "data analysis is intrinsically an empirical science" [387]. (On this basis, David Donoho argued that data science had existed for 50 years and questioned how/whether data science is really different from statistics [129]).

Data processing quickly became a critical part of the research agenda and scientific tasks, especially in statistical and mathematical domains. Typical original work on promoting data processing included *information processing* [298] and *exploratory data analysis* [388]. These works suggested that more emphasis needed to be placed on using data to suggest suitable hypotheses for testing.

Our understanding of the role of data analysis in those early years extended beyond data exploration and processing to the aspiration to "convert data into information and knowledge" [217]. The development of data processing techniques and tools has significantly motivated the proposal of the later term of "data-driven discovery" used in the first Workshop on Knowledge Discovery in Databases in 1989 [245].

Several statisticians have pushed to transform statistics to data science. For example, in 2001, an action plan was suggested in [97], in which it was suggested that the technical areas of statistics should be expanded into data science.

Prior to data science being seriously adopted in multiple disciplines, as it is today, a major analytics topic in statistics was *descriptive analytics* (also called *descriptive statistics* in the statistics community) [373]. *Descriptive analytics* quantitatively summarizes and/or describes the characteristics and measurements of a data sample or data set. Today, descriptive analytics forms the foundation for the default analytical tasks and tools in typical data analysis projects and systems.

More than 20 years after this thriving period of descriptive analytics, the desire to convert data to information and knowledge fostered the origin of the currently popular community of the ACM SIGKDD conference, specifically the first workshop on Knowledge Discovery in Databases (KDD for short) with IJCAI'1989 [245], in which "data-driven discovery" was adopted as one of three themes of the workshop.

Since the establishment of KDD, key terms such as "data mining", "knowledge discovery" [161] and *data analytics* [339] have been increasingly recognized not only in IT but also in other areas and disciplines. *Data mining* (or *knowledge discovery*) denotes the technologies and processes of discovering hidden and interesting knowledge from data.

The concept of *machine learning* was probably firstly coined by Arthur Samuel at IBM who created a checkers-playing program and defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" [187, 447].

In the history of the development of the data science community, several other major data-driven discovery-focused conferences were established in addition to the establishment of the KDD workshop in 1989. Of particular importance were the International Conference on Machine Learning (ICML) in 1980, and the Neural Information Processing Symposium (NIPS) in 1987. Since then, many regional and international conferences and workshops on data analysis, data mining, and machine learning have been created, ostensibly making this the fastest growing and most popular computer science community.

Today, in addition to well-recognized events like KDD, ICML, NIPS and JSM, many regional and international conferences and workshops on data analysis and learning have been conceived. The latest development is the creation of global and regional conferences on data science, especially the IEEE International Conference on Data Science and Advanced Analytics [135]. Data Science and Advanced Analytics has received joint support from IEEE, ACM and the American Statistical Association, in addition to industry sponsorship. These efforts have contributed to making data science the fastest growing and most popular element in computing, statistics and interdisciplinary communities.

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

The value of data analysis and data analytics has been increasingly recognized by business and management. As a result, analytics has become more data characteristics-based, business-oriented [259], problem-specific, and domain-driven [77]. Data analysis and data analytics now extend to a variety of data and domain-specific analytical tasks, such as business analytics, risk analytics, behavior analytics [74], social analytics, and web analytics. These various types of analytics are generally termed "X-analytics". Today, data analytics has become the keystone of data science.

Figure 1.2 summarizes the data science journey by capturing the representative moments and major aspects of disciplinary development, government initiatives, scientific agendas, typical socio-economic events, and education in the evolution of data science.

In Sect. 6.5.3, we discuss the evolution from processing and analysis to the broad and deep analytics of data science. Figure 6.5 demonstrates the evolutionary path from analysis to analytics and data science.

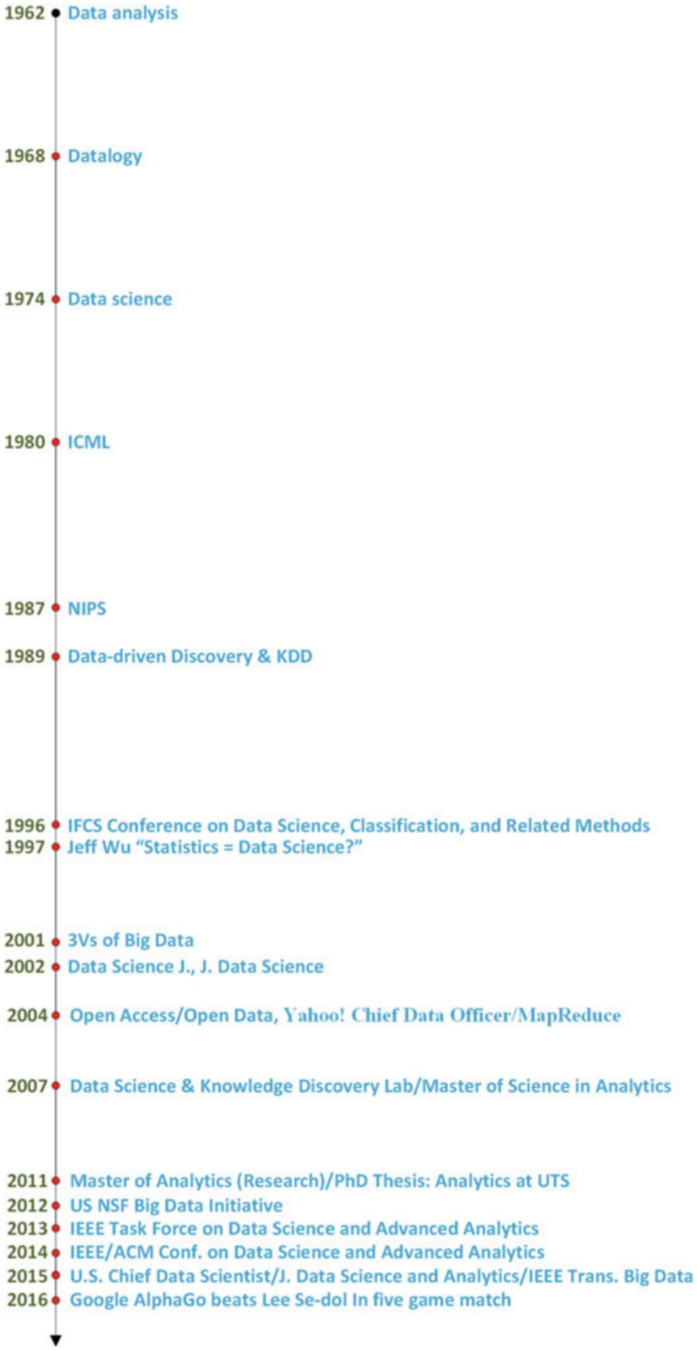

1962 Data analysis
1968 Datalogy
1974 Data science
1980 ICML
1987 NIPS
1989 Data-driven Discovery & KDD
1996 IFCS Conference on Data Science, Classification, and Related Methods
1997 Jeff Wu "Statistics = Data Science?"
2001 3Vs of Big Data
2002 Data Science J., J. Data Science
2004 Open Access/Open Data, Yahoo! Chief Data Officer/MapReduce
2007 Data Science & Knowledge Discovery Lab/Master of Science in Analytics
2011 Master of Analytics (Research)/PhD Thesis: Analytics at UTS
2012 US NSF Big Data Initiative
2013 IEEE Task Force on Data Science and Advanced Analytics
2014 IEEE/ACM Conf. on Data Science and Advanced Analytics
2015 U.S. Chief Data Scientist/J. Data Science and Analytics/IEEE Trans. Big Data
2016 Google AlphaGo beats Lee Se-dol In five game match

Fig. 1.2 Timeline of the data science journey

1.3.1 New-Generation Data Products and Economy

The disciplinary paradigm shift and technological transformation enables the innovation and industrialization of new-generation technical and practical data products and data economy.

These new-generation data products and new data economy emerge in many technical areas including data creation and quantification, acquisition, preparation and processing, sharing and storage, backup and recovery, retrieval, transport, messaging and communication, management, and governance. The dominant areas are probably the generation of new data services, such as cross-media recommender systems and cross-market financial products, as well as new data products and data systems for in-depth understanding of complex business problems that cannot be handled by existing data-driven reporting, analytics, visualization, and decision support, such as a trustful global online market supporting e-commerce of any product by anyone in any country, cross-organization data integration and analytical tools, and autonomous algorithms and automated discovery.

Another important innovation lies in the generation of domain-specific data products (including systems, applications, and services). This is typically highlighted by social media websites such as Twitter and Facebook, mobile health service and recommendation applications, online property pricing valuation and recommendation, tourism itinerary planning and booking recommendation, and personalized behavior insight understanding and treatment strategy planning.

Existing data-driven design, technologies and systems are significantly challenged by real-world human needs, which are typically intent-driven, mental, personalized, and subjective. This is reflected in online queries, preferences and demand in recommendation, online shopping and social networking. New technological innovation has to cater for these fundamental needs in the next generation of artificial intelligence and intelligent systems.

In the data and analytics areas, innovative data products, data services, and data systems may be generated in the following typical transformations:

- from a core business-driven economy to a digital and data economy;
- from closed organizations to open operations and governments;
- from traditional e-commerce to data-enabled online business;
- from landline telecommunication services to smart phone-based service mixtures that combine telecommunication and Web-based e-payment, messaging, and entertainment;
- from the Internet to mobile network and social media-mixed services; and
- from objective (object-based) businesses to subjective (intent, sentiment, personality, etc.) services.

Extended discussion on data products can be found in Sect. 2.8 and on data economy and industrialization in Chap. 8.

1.4 Data-Empowered Landscape

The disciplinary paradigm shift, technological transformation, and production of new-generation data product are driven by core data power. Core *data power* includes data-enabled opportunities, data-related ubiquitous factors, and various complexities and intelligences embedded in data-oriented transformation, production and products.

1.4.1 Data Power

Data power refers to the facilities, contributions, values, and opportunities that can be enabled by data. Data power may be reflected in different ways for different purposes. Typically, data power can be instantiated as scientific, technical, economic, cultural, social, military, political, security-oriented, and professional power.

Examples of *scientific data power* are the theoretical breakthroughs in data science research, such as new theories for learning non-IID (non-independent and identically distributed) data and new architectures for deeply representing rich hidden relations in data. Other opportunities include data-driven scientific discovery in areas that have never been explored, or that have never been possible, such as the identification of new planets and activities based on observable universal data.

Technical data power is currently widely represented by the invention of new data technologies for processing, analyzing, visualizing, and presenting complex data, such as Spark technology and Cloudera technology. Technical data power will be epitomized by novel and effective data products, data services, and data solutions that extend beyond the traditional landscape and thinking; for example, biomedical devices that can communicate with patients and understand a patient's personality and requirements.

Economic data power is reflected by the data economy and new data designs and products. The economic value of data is implemented by data-enabled industrialization, industry transformation, and productivity lift. This may lead to the development of new services, businesses, business models, and economic and commercial opportunities. It will result in smarter decision-making, more efficient markets, more personalized automated services, and best practice optimization.

Social data power is typically evidenced by social media business and social networking, which will extend into every part of our social life and society. This power creates a virtual social society in Internet and mobile network-based infrastructure that is parallel to our physical social society. The interactions and synergy between virtual society and physical society are changing our social and interpersonal relationships and lifestyle, as well as our modes of study and work. The fusion between these two worlds is significant, triggering their co-evolution and the emergence of new societal and social forms, including the way we live.

Cultural data power is progressively embodied in social data power, cultural change, and the promotion and integration of cross-cultural interactions and development. Cultural data power is also reflected in the quantification and comparison of various historical cultures, enabling global cross-culture fusion and evolution.

Military data power is deeply reflected in data-enabled and data-empowered military thinking, devices, systems, services, and methodologies. Modern military theories, systems and decisions have essentially been built on—and rely on—data. A typical example is the design of Worldwide Military Command and Control Systems [185] and the Global Command and Control System [329]. Military areas will lead data innovation, especially in fusing multiple military, professional and public systems and repositories, and developing integral detection, analysis, intervention, and weapons systems for globalized decision-making and action.

Political data power refers to the values and impact of data on politics. Political impact is reflected in data-driven evidence-based decision making, the optimization of existing policies, evidence-based informed policy-making, and optimal government services and service objectives. Significant political and governmental challenges, such as increasingly complex cross-agency decision-making and globalization-based national strategic planning, will have to rely on data fusion and deep analytics.

Security-oriented data power assures the compliance of data products by enabling the security of products and the development of data security products for more secure networks, systems, services, and devices. Secure data products, user environments, operation, data residency and sovereignty can significantly benefit from data-driven security research and innovation, complementing the traditional scope of security on infrastructure, cryptography, and protocols.

The various aspects of data power illustrated above are relative, meaning that they can be either positive or negative, depending on what drives the design, how such power is generated, and how it is utilized.

Positive data power refers to the positive value and impact that can be engendered by data. For example, algorithmic trading can identify high frequency trading strategies which can be applied to trading to increase profit.

Negative data power refers to the negative value and impact created by data. The algorithmic trading in the above example could also be used for negative purposes; for example, to manipulate high frequency trading that will result in increased personal benefit but will harm market integrity and efficiency. In this case, risk management strategies for market surveillance need to be developed to detect, predict and prevent harmful algorithmic trading.

More broadly, data power may be underused, overused or misused. Underused data power results in less competitive advantage for the data owner, e.g., the noncompetitive positioning of a company when data power is not effectively and fully utilized. Strategies, thought leadership, plans, approaches and personnel that can take full advantage of data power are necessary. In contrast, overused and misused data power may generate misleading or even unlawful outcomes and impact. Assessment, prediction, prevention and intervention strategies, systems and capabilities must be developed to detect and manage negative data power.

How the power of data is recognized, fulfilled and valuated may determine the strategic position, competitive advantage, tools, and development of a data-intensive organization. The level at which this is conducted is critical for countries, enterprises and individuals. With the emergence of many new companies built on recognizing and utilizing specific data power, as is evident in the increasingly growing big data landscape, entities that ignore the strategic value of data power may significantly lag behind and be disadvantaged. The imbalanced development of a country, an enterprise, or an individual may be the result of ineffective and/or inefficient recognition of data power, and the consequent vision, and actions of achieving data power. Competing in the fourth revolution in data-driven science, technology and economy, a fundamental and strategic matter is to study data power, and create corresponding early-bird vision, strategies, initiatives, and actions to take advantage of data power from political, scientific, technological, economic, educational, and societal perspectives.

1.4.2 Data-Oriented Forces

Ubiquitous data-oriented driving forces can be seen from the viewpoint of both high- and low-level vision and mission, given the prevailing data, behavior, complexity, intelligence, service and opportunity perspectives.

Vision and mission determine the big picture and strategic objectives, and the view of what data will satisfy organizational strategic needs and requirements, and how. Strategic, forward-looking, long-term and big picture thinking is required. This is often challenging, as few people have the training, capability or mindset for such purposes.

Technical and pragmatic data driving forces directly involve data-oriented elements: data, behavior, complexity, intelligence, service and future.

- *Data* is ubiquitous, and includes historical, real-time, and future data;
- *Behavior* is ubiquitous, and bridges the gaps between the physical world and the data world;
- *Complexity* is ubiquitous, and involves the type and extent of complexity that differentiates one data system from another;
- *Intelligence* is ubiquitous, and is embedded in a data system;
- *Service* is ubiquitous, and is present in multiple forms and domains; and
- *Future* is unlimited with ubiquitous opportunities, because data enables enormous opportunity.

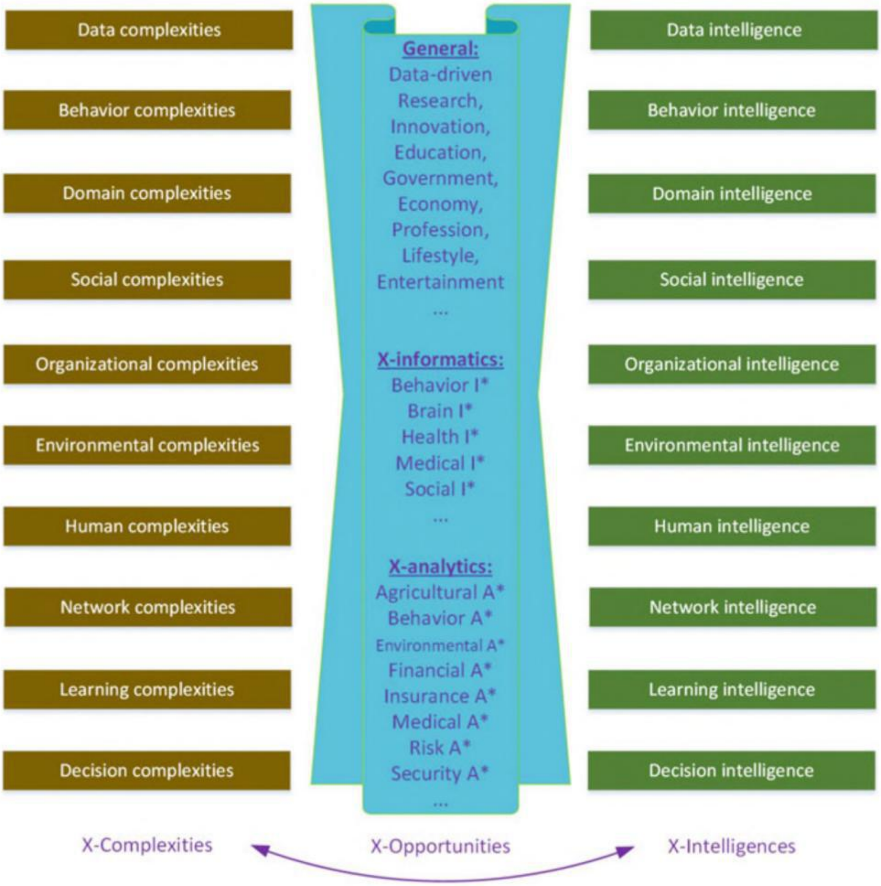


Fig. 1.3 The new X-generations: X-complexities, X-intelligence, and X-opportunities

1.5 New X-Generations

As a result of data business achieving a level of importance that is comparable to traditional, physical business, the world is experiencing a revolutionary migration of complexities, intelligences, and opportunities to their new X-generations. X-generations are embodied in both (1) fundamental driving areas: X-complexities (see Sect. 1.5.1) to be addressed and X-intelligence (see Sect. 1.5.2) to be involved or created, and (2) strategic potential: X-opportunities, such as X-analytics (see Sect. 7.6) and X-informatics (see Sect. 1.5.3) to be generated.

Figure 1.3 illustrates the aspects and perspectives related to X-complexities, X-intelligence, and X-opportunities, which are briefly explained below. Other X-generations are X-analytics and X-informatics, as discussed in Sect. 7.6.

1.5.1 X-Complexities

A data science problem is a complex system [62, 294] that has a variety of intrinsic system complexities. The study of data science has to tackle multiple complexities which have not been addressed or addressed well. This new generation of data-driven science, innovation and business relies on the exploration and utilization of complexities that have not previously been well characterized and addressed, if at all.

In complex data science problems, *X-complexities* [62, 64] refers to diverse, widespread complexities that may be embedded in data, behavior, domain, societal aspects, organizational matters, environment, human involvement, network, and learning and decision-making. These complexities are represented or reflected by such factors as those given below.

- *Data complexity* Comprehensive data circumstances and characteristics;
- *Behavior complexity* Individual and group activities, evolution, utility, impact, and change;
- *Domain complexity* Domain factors, processes, norms, policies, knowledge, and the engagement of domain experts in problem solving;
- *Social complexity* Social networking, community formation and divergence, sentiment, the dissemination of opinion and influence, and other social issues such as trust and security;
- *Environment complexity* Contextual factors, interactions with systems, changes, and uncertainty;
- *Learning complexity* Including the development of appropriate methodologies, frameworks, processes, models and algorithms, and theoretical foundation and explanation;
- *Human complexity* The involvement and roles of human beings, human intelligence and expert knowledge in data science problems, systems and problem-solving processes; and
- *Decision-making complexity* Methods and forms of deliverables, communications and decision-making actions.

More discussion about X-complexities from the data science challenge perspective will be conducted in Sect. 4.2.

1.5.2 X-Intelligence

A complex system is usually embedded with mixed intelligence, and in the data world, data systems are intelligence-based systems. In a complex data science problem, ubiquitous intelligence, called *X-intelligence* [62, 64], is often demonstrated.

X-intelligence is embedded with X-complexities and consists of data intelligence, behavior intelligence, domain intelligence, human intelligence, network intelligence, organizational intelligence, and environmental intelligence, which are briefly discussed below.

- *Data intelligence* highlights the interesting information, insights, and stories hidden in data about business problems and driving forces.
- *Behavior intelligence* demonstrates the insights of activities, processes, dynamics, impact, and the trust of individual and group behaviors by humans and action-oriented organisms.
- *Domain intelligence* includes domain values and insights that emerge from domain factors, knowledge, meta-knowledge, and other domain-specific resources.
- *Human intelligence* includes contributions made by the empirical knowledge, beliefs, intentions, expectations, critical thinking, and imaginary thinking of human individuals 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 in other intelligences specific to the underlying domain, organization, society, and actors.

X-intelligences in a data science system are mixed. They interact with each other and may not be easily decomposed. A good data product must effectively represent, incorporate and synergize core aspects of X-intelligence that play a fundamental role in system dynamics and problem-solving processes and systems.

More discussion about X-intelligences is available in Chap. 1 in book [68].

1.5.3 X-Opportunities

Our experience and literature review also confirm that data science enables unimagined general and specific opportunities, called *X-opportunities*, for

- *new research*: i.e., “what I can do now but could not do before”;
- *better innovation*: i.e., “what I could not do better before but I can do well now.”
- *new business*: i.e., “I can make money out of data.”

X-opportunities from data may be general or specific. General X-opportunities are enormous and overwhelming. They extend from research, innovation and education to new professions, new ways of operating government, and new economy. In fact, as new models and systems of data-driven economy and research findings emerge, it is a matter of how our imagination can perceive these opportunities. New data products and services emerge as a result of identifying new data-driven business models and opportunities.

We highlight the directions for creative data-driven opportunities below.

- *Research*, such as inventing data-focused breakthrough theories and technologies;
- *Innovation*, such as developing cutting-edge data-based intelligent services, systems, and tools;
- *Education*, such as innovating data-oriented courses and training;
- *Government*, such as enabling data-driven evidence-based government decision-making and objective planning and execution;
- *Economy*, 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.

Data-driven opportunities are unlimited, especially in the scenario in which "I do not know what I do not know". Simply by recognizing some of the potential opportunities, the world could be incrementally or significantly changed. To have the capacity to recognize more data-driven opportunities, we need data science thinking.[1] Being creative and critical is important for detecting new opportunities.[2]

X-opportunities may be specified in terms of particular aspects, problems, and purposes. *X-informatics*, which refers to the creation and application of informatics for specific domain problems, is one instance. Another instance is *X-analytics*, which refers to the various opportunities discoverable by applying and conducting analytics on domain-particular data.

Examples of X-informatics are behavior informatics, brain informatics, health informatics, medical informatics, and social informatics. More discussion about informatics for data science can be found in Sect. 6.4.2.

Instances of X-analytics are agricultural analytics, behavior analytics, disaster analytics, environmental analytics, financial analytics, insurance analytics, risk analytics, transport analytics, and security analytics. More discussion about X-analytics is available in Chap. 3 in book [67].

1.6 The Interest Trends

Prior to the prevalence of big data, data analysis, data analytics, and data science were attracting growing attention from several communities, in particular statistics. In recent years, big data analytics, data science, and advanced analytics have become

[1] See more discussion about data science thinking in Chap. 3.

[2] Refer to Sect. 3.2.2 and in particular Sect. 3.2.2.3 for more discussion about creative and critical thinking in data science.

increasingly popular in not only the broad IT area but also in other disciplines and domains.

According to Google Trends [193], the online search interest over time in "data science" is similar to the interest in "data analytics", but is 50–100% less than the interest in "big data". However, the historical search interest in data science and analytics is roughly double the interest shown in big data about 10 years ago. Compared to the smooth growth of interest in data science and analytics, the interest in big data has experienced a more rapid increase since 2012. When we googled "data science", 83.8M records were returned, compared to 365M on "big data", and 81.8M on "data analytics".[3]

Although they do not reflect the full picture, the Google search results over the last 10 years, shown in Fig. 1.4, indicate that:

- Data science, data analysis, and data analytics have much richer histories and stronger disciplinary foundations than big data.
- The significant boom in big data has been fundamentally business-related, while data science has been highly linked with research and innovation.
- Data analysis has always been a top concern, although search interest has been flattened and diversified into other hot topics, including big data, data science and data analytics.
- Interestingly, the word "advanced analytics" has received much less attention than all other terms, reflecting the fact that knowledge of, and interest in, more general terms like data analytics is greater than it is for more specific terms such as advanced analytics.
- Compared to 10 years ago, scrutiny of the search trends in the past 4 years would find that big data saw significantly increasing interest from 2012 to 2015, followed by a period of less movement; however, the interest in data science and data analytics has consistently increased, although it has grown at a much lower rate (some one third of big data). Data analysis has maintained a relatively stable attraction to searchers during these 10 years.

1.7 Major Data Strategies by Governments

Governments play the driving role in promoting and operationalizing data science innovation, big data technology development, data industrialization and data economy formation. This section summarizes the representative data strategies and initiatives established by global governments and the United Nations [63].

[3] Note, these figures were collected on 15 November 2016.

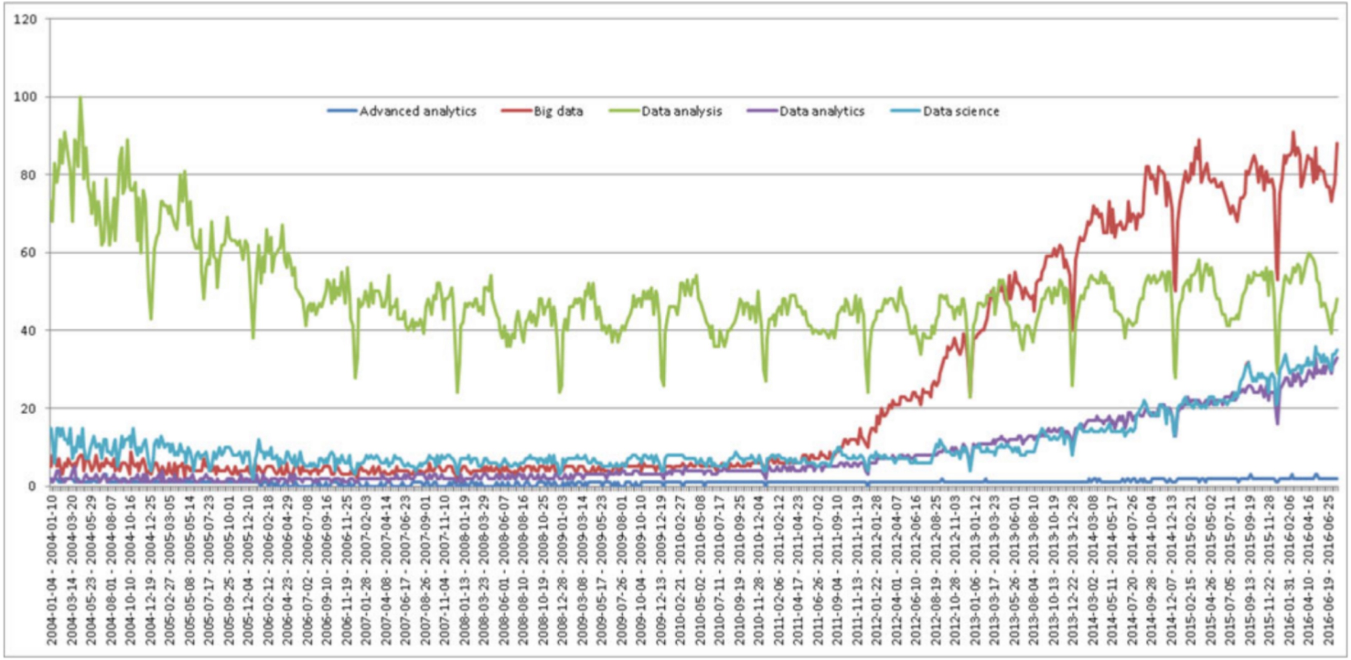


Fig. 1.4 Online search interest trends on data science-related key terms by Google. Note: data was collected on 15 November 2016

Table 1.3 Government initiatives in big data and data science

Government	Representative initiatives
Australia	Public Service Big Data Strategy [399], Whole-of-Government Centre of Excellence on Data Analytics [17]
Canada	Capitalizing on Big Data [50]
China	Big Data Guideline [196], China Computer Federation Task Force on Big Data [85], China National Science Foundation big data program [99]
EU	Data-driven Economy [102], European Commission Horizon 2020 Big Data Private Public Partnership [214]
UK	UK's Big Data and Energy Efficient Computing [395]
UN	UN Global Pulse Project [399]
US	US Big Data Research Initiative [407], Interagency Working Group on Digital Data [107], DARPA's XDATA Program [115], USA NSF Big Data Research Fund [407]

1.7.1 Governmental Data Initiatives

To effectively understand and utilize everywhere data, data DNA and its potential, increasing numbers of regional and global government data initiatives [356] are being introduced at different levels and on different scales in this age of big data and data science to promote data science research, innovation, funding support, policy making, industrialization, and economy.

Table 1.3 summarizes the major initiatives of several countries and regions.

1.7.2 Australian Initiatives

The Australian Government has published several papers on big data and analytics. In the Big Data Strategy—Issues Paper [4], major issues related to big data were discussed. The Australian Public Service Big Data Strategy paper set up its goal to address big data strategy issues and to "provide an opportunity to consider the range of opportunities presented to agencies in relation to the use of big data, and the emerging tools that allow us to better appreciate what it tells us, in the context of the potential concerns that this might raise" [4].

Australia's whole-of-government Centre of Excellence in Data Analytics [17] coordinates relevant government activities. It encourages respective government agencies and departments to think about, promote and accept data-driven approaches for government policy optimization, service improvement and cross-government operations.

As part of a critical government strategy, the Australian Research Council has granted approval to the Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers [1] to conduct research on big data-based mathematical and statistical foundations.

Another recent effort made by the Australian Government was the establishment of Data61 [116], which consolidated data-related human resources in the former National Information and Communications Research Centre of Australia (NICTA) [308] and Commonwealth Scientific and Industrial Research Organisation (CSIRO) and aims to achieve a unified platform for data research and innovation, engagement with industry, government and academia, and software development.

1.7.3 Chinese Initiatives

The Chinese Government treats big data as an essential constituent in its government strategic plan, innovation strategies, and economic transformation. Big data has been an increasingly topical concept in central, state and city-based Chinese government policies, plans and activities. Increasing numbers of facilities, funding, and initiatives have been committed to this area. The Chinese Government is particularly interested in big data-driven innovation and economy.

Several guidance documents and plans have been issued by the central Chinese Government. China's Guidelines [196], for example, are aimed at boosting the development of big data research and applications, to "set up an overall coordination mechanism for big data development and application, speed up the establishment of relevant rules, and encourage cooperation between the government, enterprises and institutions."

China has also set up a national strategic plan for the IoT and big data [196]. Research and commercial-ready funding have been committed to national key research projects and key national labs for big data research, innovation and industrialization.

Many states and cities in China, such as Beijing [195], have launched national big data strategies and action plans for big data and cloud computing [3, 84]. A very early example in China was the Luoyang City-sponsored consulting project in 2011, for which we developed a strategic plan for the City's industrial transformation to a "data industry" [56].

Several data technology parks have been established in China. The Ministry of Education in China recently approved about 30 applications of opening undergraduate courses in data science and big data technology. China has experienced very fast conceptual shift in recent years, from the Internet of Things to cloud computing, and then to big data and now artificial intelligence.

1.7.4 European Initiatives

The European Union (EU) and its member countries consider data-driven economy as a new transformation opportunity and strategy. Several initiatives have been made, and some examples are given below.

The European Commission's (EC) EU communication Towards a Thriving Data-driven Economy [102] is "an action plan to bring about the data-driven economy of the future", which outlines "a new strategy on Big Data, supporting and accelerating the transition towards a data-driven economy in Europe. The data-driven economy will stimulate research and innovation on data while leading to more business opportunities and an increased availability of knowledge and capital, in particular for SMEs, across Europe."

In 2015, the European Data Science Academy, EDSA [154] was formed. EDSA will analyze the skillsets for data analysts, develop modular and adaptable curricula, and deliver multiplatform and multilingual training curricula.

Member countries in the EU establish their own big data strategies and initiatives. For example, the European Commission published a communication entitled "Towards a thriving data-driven economy" [102] aims to address the big data challenge by "sketching the features of the European data-driven economy of the future and setting out some operational conclusions to support and accelerate the transition towards it." The directive also sets out current and future activities in the field of cloud computing.

1.7.5 United States' Initiatives

The United States (US) has played the leadership role in the global promotion of big data and data economy. The US Government has initiated a government strategy and funding support for big data research and industrialization, and several of its initiatives are highlighted below.

Big Data Research Initiative [407] is directed toward "supporting the fundamental science and underlying infrastructure enabling the big data revolution." In 2005, the US National Science Board set the goal that it "should act to develop and mature the career path for data scientists" in its report "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" [310].

In 2009, the Committee on Science of the National Science and Technology Council formed an Interagency Working Group on Digital Data which published a report [107] outlining the strategy to "create a comprehensive framework of transparent, evolvable, extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data", which "will serve as a driving force for American leadership in science and in a competitive, global information society."

In addition, the Defence Advanced Research Projects Agency (DARPA) launched its XDATA Program [115], which aims to develop computational techniques and software tools for processing and analyzing large, imperfect and incomplete data. In 2012, the National Institute of Standards and Technology (NIST) introduced a new data science initiative [130], and in 2013, the US National Consortium for Data Science was established [405].

Many US vendors certainly drive the development of big data economy, as highlighted by the significant growth of new data-focused companies such as Google, Facebook, Spark, and Rapidminer, and traditional IT companies such as Microsoft and Oracle.

1.7.6 Other Governmental Initiatives

The United Nations (UN) Global Pulse Project is "a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate the discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action." [399]

Most countries have made plans, or at least are making plans, to promote a data-driven economy. For example, Canada's policy framework Capitalizing on Big Data [50] aims at "establishing a culture of stewardship . . . coordination of stakeholder engagement . . . developing capacity and future funding parameters."

The United Kingdom's Big Data and Energy Efficient Computing initiative funded by Research Councils UK [395] aims to "create a foundation where researchers, users and industry can work together to create enhanced opportunities for scientific discovery and development."

Without any doubt, an increasing number of government efforts and investment will be made in big data, cloud computing, data science and artificial intelligence.

1.8 The Scientific Agenda for Data Science

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

1.8.1 The Scientific Agenda by Governments

The original scientific agenda of data science has been driven by both government initiatives and academic recommendations, building on the strong promotion of converting statistics to data science, and blending statistics with computing science in the statistics community [97, 128, 164, 197, 203, 205, 231, 467].

Today, many regional and global initiatives have been taken in data science research, disciplinary development and education, creating a data power-enabled strategic agenda for the data era. Several examples are given below.

- In Australia, a Group of Eight (Go8) report [45] suggested the incorporation of data as a keystone in K-12 education through statistics and science by such methods as creating data games for children.
- In China, the Ministry of Science and Technology very recently announced the establishment of national key labs in big data research as part of a strategic national agenda [98].
- In the EU, the High Level Steering Group (HLSG) report of the Digital Agenda for Europe "Riding the Wave" [211] and the Research Data Alliance (RDA) report "The Data Harvest" [212], urged the European Commission to implement the vision of creating "scientific e-infrastructure that supports seamless access, use, re-use, and trust of data" and to foster the development of data science university programs and discipline.
- In the US, a National Science Board report [310] recommended that the National Science Foundation (NSF) "should evaluate in an integrated way the impact of the full portfolio of programs of outreach to students and citizens of all ages that are 'or could be' implemented through digital data collections." Different roles and responsibilities were discussed for individuals and institutions, including data authors, users, managers and scientists as well as funding agencies. The report [107] from the US Committee on Science of the National Science and Technology Council suggested the development of necessary knowledge and skill sets by initiating new educational programs and curricula, such as "some new specializations in data tools, infrastructures, sciences, and management."

1.8.2 Data Science Research Initiatives

An increasing number of research streams, strengths and focused projects have been announced in major countries and regions. Examples include:

- The US NSF Big Data Research Fund [407],
- The European Commission Horizon 2020 Big Data Private Public Partnership [102, 214], and
- The China NSF big data special fund [99].

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

Many universities and institutions have either established or are creating research centers or institutes in data science, analytics, big data, cloud computing, and IoT. In Australia, for example, the author created the first data science lab: the Data Science and Knowledge Discovery Lab at UTS in 2007 [141], and the first Australian institute: the Advanced Analytics Institute [4, 409] in 2011 which implements the RED model of Research, Education and Development (RED) of big data analytics for many major government and business organizations. In the US, top universities have worked on building data science initiatives, such as the Institute for Advanced Analytics at North Carolina State University in 2007 [302], the Stanford Data Science Initiatives in 2014 [370], and the Data Science Initiatives at University of Michigan in 2015 [398].

1.9 Summary

As we stepped into the twenty-first century, significant opportunities and a fundamental revolution in economy and science were driven by the worldwide increase in big data, although few people recognized or anticipated the compelling changes that would result.

It is amazing to think that Yahoo created the important role of "Chief Data Officer" in early 2000, and that the term "data science" was created as long as 50 years ago. It is big vision that has driven this world revolution, yet although we are lucky to live at this exciting time, our limited imagination means that we have been slow to embrace and guide a different age: the data era.

Although this new era has been termed "the era of big data", "the age of analytics", and "the data science era", the data era effectively extends far beyond data science, big data, and advanced analytics. It has the capacity to fundamentally change our life, society, way of living, entertainment, and of course the way we work, study and do business. The data era opens the door to a new age of data-driven science and economy.

This chapter has somewhat limited the scope of the data era to one that is data science-based, but our thinking should exceed this limitation. The content of this book thus goes beyond the scientific component in several chapters to embrace concepts about the data economy, data industrialization, and data professions. It is hoped that this will complement the introduction in this chapter and compel the reader's thinking to another age: that of data-driven future science, economy, and society.

This section discusses these relevant and important concepts.

It is challenging to quantify these concepts and their differences. An empirical understanding of the conceptualization has been provided in the so-called DIKW Pyramid [342, 425], its earlier variation "signal, message, information, and knowledge" [38], and various arguments and refinements of these conceptualizations [62].

These concepts, including intelligence, capture the different existing forms and progressive representations of objects (or entities) in a cognitive processing and production system.

Data, represents discrete or objective *facts*, *signals* (from sensors, which may be subjective or objective), or *symbols* (or signs) about an object (a physical or virtual entity or event). Data is at the lowest level of cognitive systems, can be subjective or objective, can be with or without meaning, and has a value. An example of data is "8 years old" or "young."

Information, represents a description of relevant data (objects) in an organized way, for a certain purpose, or having a certain meaning. Information can be structural (organized) or functional (purposeful), subjective (relevant to an intent) or objective (fact-based). For example, "Sabrina is 8 years old" is a piece of information which describes a structured relationship between two objects, "Sabrina" and "8 years old". Another example is "Sabrina is young."

Knowledge, represents the form of processed information in terms of an information mixture, procedural actions, or propositional rules. Knowledge can be subjective or objective, known or unknown, actionable or not, and reasonable or not. Examples of processed, procedural and propositional knowledge are "Year 3 students are mostly 8 years old", "Tea and medicine are not supposed to be taken at the same time," and "All 8 year old children should go to school."

Intelligence, representing the ability to inform, think, reason, infer, or process information and knowledge. Intelligence is either inbuilt or can be improved through learning, processing or enhancement. Intelligence can be high or low, hierarchical, general or specific. Examples of intelligence are: "Sabrina is probably in Year 3" (reasoning outcome based on the fact that she is 8 years old and a child of that age is usually at school) or "Sabrina is intelligent" (based on the information that she has always attained high marks at school.)

Wisdom, represents a high-level principle, which is the cognitive and thinking output of information processing, knowledge management, or simply inspiration gained through experiences or thinking. Wisdom indicates the superior ability, meta-knowledge, understanding, application, judgment, or decision-making inherent in knowing or determining the right thing to do at the right time for the right purpose. Wisdom can be non-material, unique, personal, intuitive, or mentally-inspired. Compared to knowledge, wisdom is timeless, comprehensive, general, and sentimental, being passed down in histories and cultures in the form of common sayings, quotations, or philosophical phrases. Examples of wisdom are "A young idler, an old beggar," and "The child is father of the man."

It is difficult to generate a simple framework to show the difference between data, information, knowledge, intelligence, and wisdom. Figure 2.1 illustrates the relationships between them and the path of progression from data to wisdom. In

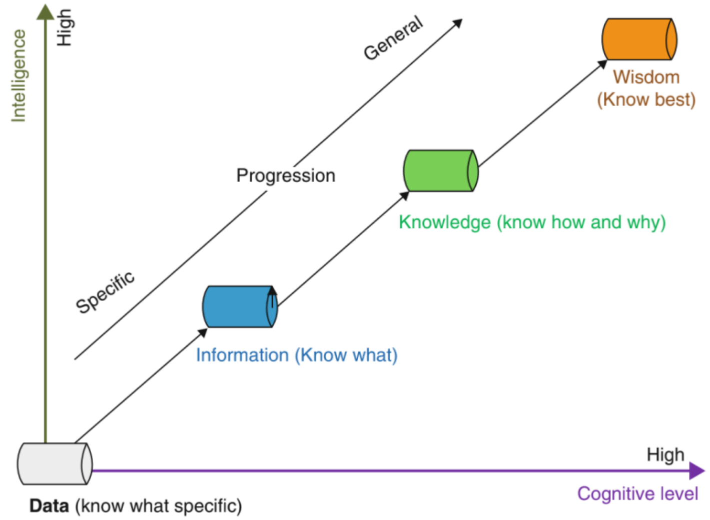


Fig. 2.1 Data-to-information-to-knowledge-to-intelligence-to-wisdom cognitive progression. Note: X-axis: the increase in cognitive level; Y-axis: the increase in intelligence

this framework, the data-to-wisdom path is a specific-to-general progressive journey according to the increase in cognitive level and intelligence.

- Data is about *the aspect of a subject*;
- Information describes *what is known about a subject*, i.e., *know what* (including *know who*, *know when*, *know where*, etc.) in relation to data;
- Knowledge concerns the *know how* and *know why* (or *why is*) about information;
- Wisdom is the intelligence to *know best* about *how to act* (or *why to act*) on the basis of a usually widely validated ability or understanding.

During the production and cognitive processing procedure, intelligence plays an enabling role for both progression and production.

In addition, as discussed in Sect. 4.2 about X-complexities and Sect. 4.3 about X-intelligence, the progression of data-to-information-to-knowledge-to-wisdom needs to involve and handle relevant complexities and intelligences.

2.4 Data DNA

2.4.1 What Is Data DNA

In biology, DNA is a molecule that carries genetic instructions that are uniquely valuable to the biological development, functioning and reproduction of humans and all living organisms.

As a result of data quantification, data is everywhere, and it is present in the public Internet; the Internet of Things (IoT); sensor networks; sociocultural, economic and geographical repositories; and quantified personalized sensors, including mobile, social, living, entertaining, and emotional sources. These form 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 2.1 (Data DNA) *Data DNA* is the datalogical "molecule" of data, consisting of fundamental and generic constituents: entity (E), property (P), behavior (B), and relationship (R).

Here, "datalogical" means that data DNA plays a similar role in data organisms as biological DNA plays in living organisms. The four elements in data DNA, namely behavior, entity, relationship and property (BERP), represent diverse but fundamental aspects in data. *Entity* can be an object, instance, human, organization, system, or part of a subsystem, or environment. *Property* refers to the attributes that describe an entity. *Behavior* refers to the activities and dynamics of an entity or a collection of entities. *Relationship* corresponds to entity interactions and property interactions, including property value interactions.

2.4.2 Data DNA Functionalities

Entity, property, behavior and relationship have 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* (the 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 [417] and non-personal data organisms, and in the generation, development, functioning, management, analysis, and use of all data-based applications and systems.

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

In complex data, data DNA is embedded within various X-complexities (see detailed discussion in Sect. 1.5.1 and in [64] and [62]) and ubiquitous X-intelligence (more details in Sect. 1.5.2 and in [64] and [62]) in a data organism. This makes data rich in content, characteristics, semantics, and value, but challenging in acquisition, preparation, presentation, analysis, and interpretation.

2.5 Data Science Views

In this section, the different views of data science are discussed to create a picture of what makes data science a new science.

2.5.1 The Data Science View in Statistics

Statisticians have had much to say about data science, since it is they who actually created the term "data science" and promoted the upgrading of statistics to data science as a broader discipline.

Typical statistical views of data science can be reflected in the following arguments and recommendations.

In 1997, Jeff Wu questioned whether "Statistics = Data Science?". He suggested that statistics should be renamed "data science" and statisticians should be known as "data scientists" [467]. The intention was to shift the focus of statistics from "data collection, modeling, analysis, problem understanding/resolving, decision making" to future directions on "large/complex data, empirical-physical approach, representation and exploitation of knowledge".

In 2001, William S. Cleveland suggested that it would be appropriate to alter the statistics field to data science and "to enlarge the major areas of technical work of the field of statistics" by looking to computing and partnering with computer scientists [97].

Also in 2001, Leo Breiman suggested that it was necessary to "move away from exclusive dependence on data models (in statistics) and adopt a more diverse set of tools" such as algorithmic modeling, which treats the data mechanism as unknown [42].

In 2015, a statement about the role of statistics in data science was released by a number of ASA leaders [145], saying that "statistics and machine learning play a central role in data science." Many other relevant discussion is available in AMSTATNEWS [12] and IMS [473].

2.5.2 A Multidisciplinary Data Science View

In recent years, data science has been elaborated beyond statistics. This is driven by the fact that statistics cannot own data science, and the statistics community has realized the limitation of statistics-focused data science and the broader capability requirements that go beyond statistics.

A multidisciplinary view has thus been increasingly accepted not only by the statistics community, but also other disciplines, including informatics, computing and even social science. This reflects the progressive evolution of the concept and vision of data science, from statistics to informatics and computing, as well as other fields, and the interdisciplinary and cross-disciplinary nature of data science as a new science.

Intensive discussion has taken place in the research and academic communities about creating data science as an multidisciplinary academic field. As a new discipline [364], data science involves not only statistics, but also other disciplines. The concept of data science is correspondingly defined from the perspective of disciplinary and course development [470].

Although different communities may share contrasting views about what disciplines are involved in data science, statistics, informatics, and computing are three fields that are typically viewed as the keystones, making data science a new science.

In addition, some people believe a cross-disciplinary body of knowledge in data science includes informatics, computing, communication, management, and decision-making; while others treat data science as a mixture of statistics, mathematics, physics, computer science, graphic design, data mining, human-computer interaction, information visualization, and social science.

Today, there is increasing consensus that data science is inter-disciplinary, cross-disciplinary, and trans-disciplinary. We will further discuss the definition of data science from the disciplinary perspective in Sect. 2.6.2.

2.5.3 The Data-Centric View

Although there are different views or perspectives through which to define what makes data science a new science, a fundamental perspective is that *data science is data centric*. There are several aspects from which to elaborate on the data-centric view: hypothesis-free exploration, model-independent discovery, and evidence-based decision-making, to name three.

First, *hypothesis-free exploration* needs to be taken as the starting point of data understanding. There is no hypothesis before a data science task is undertaken. It is data that generates, indicates, and/or validates a new hypothesis. New hypothesis generation relies greatly on a deep understanding of the inbuilt data characteristics, complexities and intelligence of a problem and its underlying environment.

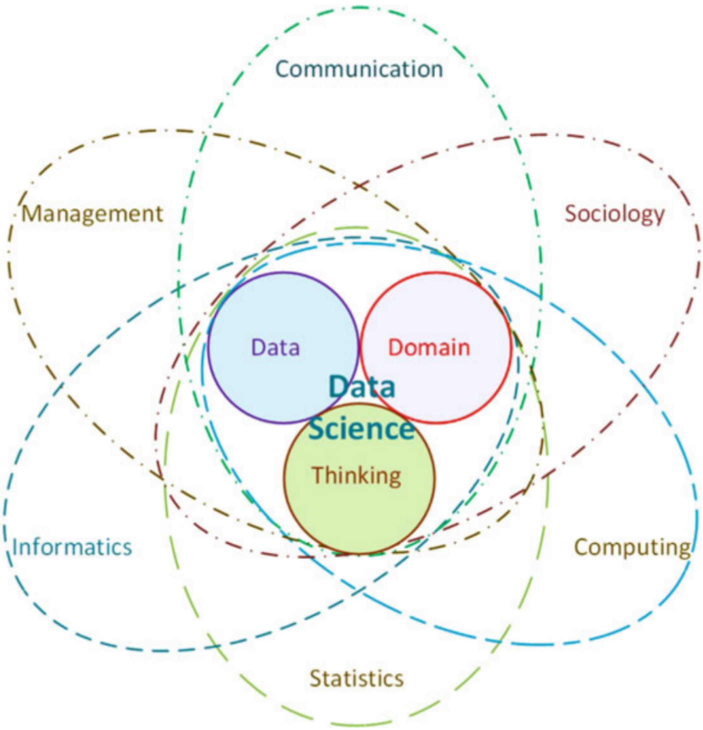


Fig. 2.2 Trans-disciplinary data science

science fields, such as behavioral data science, health data science, or even history-based data science.

2.6.3 *Process-Based Data Science Definition*

Generally speaking, *data science is the science (or study) of data* as defined in Data Science[1]. However, there are different ways of specifying what data science is; it may be object-focused, process-based, or discipline-oriented [64], as in Data Science[2]. From the data science process perspective, we offer the following definition, building on, involving and/or processing DIKIW.

Definition 2.4 (Data Science[3]) From the *DIKIW-processing* perspective, *data science* is a systematic approach to "thinking with wisdom", "understanding the domain", "managing data", "computing with data", "discovering knowledge", "communicating with stakeholders", "acting on insights", and "delivering products".

2.6.3.4 Computing with Data

Computing with data refers to how to manipulate data for what purposes and in what ways. Computing with data may consist of tasks such as feature engineering, data exploration, descriptive analysis, visualization, and data presentation.

Feature engineering consists of the understanding, selection, extraction, construction, fusion, and mining of features, which are fundamental for data use, knowledge discovery, and other data-driven learning.

Data exploration is to understand, quantify and visualize the main characteristics of data. This may be achieved by descriptive analysis, statistical methods, and visual analytics.

Descriptive analysis is typically statistical analysis. It may involve the quantitative examination, manipulation, summarization, and interpretation of data samples and data sets to discover underlying trends, changes, patterns, relationships, effects and causes.

Data visualization presents data in a visual way, i.e., in a pictorial or graphical format. Typically, charts, graphs, and interactive interfaces and dashboards are used to visualize patterns, trends, changes, and relationships hidden in data, business evolution, and problem development.

Data presentation involves the broad communication of data and data products. Typical data presentation tools and means include reports, dashboards, OLAP, and visualization.

In Sect. 6.6, more discussion about computing for data science is provided.

2.6.3.5 Discovering Knowledge

Knowledge discovery [161, 202] is a higher level of data manipulation that aims to identify hidden but interesting knowledge which discloses intrinsic insights about problems, problem evolution, causes, and effects.

Typical knowledge discovery tasks including prediction, forecasting, clustering, classification, pattern mining, and outlier detection. When knowledge discovery is applied to specific data issues, it generates respective knowledge discovery methods, tasks and outcomes. For example, climate change detection seeks to detect significant changes taking place in the climate system.

In data science, critical issues to consider in knowledge discovery are

- what knowledge is to be mined?
- is the knowledge actionable, trustful, and transformative? and
- how can the knowledge be converted to insight, intelligence and wisdom?

The *actionablity* [59, 77] of knowledge determines how useful the discovered knowledge will be for decision-making and game changing. The transformation from knowledge to intelligence and wisdom requires additional theories and tools, which could involve X-intelligence (see more in Sect. 4.3), including domain intelligence, social intelligence, organization intelligence, and human intelligence.

This is because intelligence and wisdom aims to achieve a high level and general abstraction of data values.

2.6.3.6 Communicating with Stakeholders

Communicating with stakeholders is particularly important for data science tasks. *Communicating with stakeholders* involves many important aspects, such as:

- Who are the stakeholders of a data science project?
- How can you communicate with each stakeholder?
- What is to be communicated to each stakeholder?
- What are the skills required for good communications in data science?

Stakeholders in a data science project may be of several types, at various levels: decision makers, project owners, data scientists, business analysts, business operators. They have different roles and responsibilities in an organization, and in the data science process and engineering function. Communications with each role and between roles may involve disparate objectives, channels, and content.

As data scientists are the core players in data science projects, it is critical for them to communicate with business owners and business operators about the business objectives, requirements, scope, funding models, priorities, milestones, expectations, evaluation methods, assessment criteria, implementation requirements, and deployment process of data science projects.

There are different means of initiating and undertaking communications with and between the various roles in a data science project. Executive reports are common for executives and decision makers. Business analysts may like more quantitative analytical reports with evidence and supporting justifications. Business operators may prefer dashboards and user-friendly “automated” systems.

Many skills may be required to achieve good communications. For example, summarization, abstraction, generalization, visualization, formalization, quantification, representation, and reporting may be used for different purposes on different occasions.

In Sect. 6.9, the roles and relations between communications (as a discipline) and data science are discussed.

2.6.3.7 Delivering Data Products

Data science needs to deliver outcomes, which we call *data products*. As defined in Sects. 2.8 and 1.3.1, data products refer to broad data-related deliverables.

Often data products are equivalent to the knowledge and findings discovered in a data analytical project, but this does not reflect the full picture of data science deliverables.

resolution of scientific problems, and funding and projects to support research and innovation and the evaluation, dissemination and management of scientific results. Typical activities to support open science are crowdsourcing, international collaborative open projects, and research networks.

Open science data and open access are two necessary mechanisms for enabling open science and innovation. Science data are freely available to the public. Open access [451] is a principle and mechanism for enabling free access to scientific outputs (including peer-reviewed and non-peer-reviewed results) from scientific activities through scientific dissemination channels (including journals and conferences). Such freely accessible data are archived and managed by scientists, institutions or independent organizations. The authors of the scientific outputs control the copyright of their work when it is published or republished, the integrity of their work, and the right of their work to be lawfully used and acknowledged.

Open source [456] refers to the principle, methodology and mechanisms for creating, distributing, sharing, and managing software, also called *open source software*. The source codes are made freely available to the public, hence this is also known as *free software*. Open source software is typically associated with a licensing arrangement that allows the copyright holder to decide how the software is distributed, changed and used by others, and for what purpose. Open source software requires corresponding software development infrastructure; collaborative development models, methods, platforms and specifications; and copyright, laws, agreements, and norms for certification, distribution, commercialization and licensing, change, usage rights and risk management (e.g., security).

Open review and evaluation [454] are the review and evaluation, process and management mechanisms of scientific outputs by the public or peer reviewers whose names are disclosed to the authors. The review and evaluation process, commenting activities and reports, revision and responses between authors and reviewers may be open in a public (Internet-driven) or review management system (e.g., journal review system).

Open education and training [453] refers to the online provision of educational and training admission, course-offering, resource sharing, teaching-learning servicing, and accreditation. Open education and training exceeds the limitations of awarded courses and short courses that are traditionally offered through educational and training institutions. Open education and training changes the way that scientific knowledge and capabilities are transferred to learners by providing more ad hoc, flexible, and customizable study plans, channels, scheduling, course formation, and resources. It removes the restrictions on course availability, comparison, selection and change, lecturers, study modes, scheduling, and materials that are a feature of institution-based education and training, and enables learners to make choices and advance their learning through the global, fast and flexible approach enabled by Internet-based online courses. Open education thus encourages better teaching and learning quality and performance. The open science movement has motivated the emergence of many open movements and activities in a range of scientific disciplines, scientific research processes, enabling and support facilities, and in

411. Vast: Visual analytics community (2016). URL http://vacommunity.org/HomePage
412. Veaux, R.D.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.X., Washington, T.M., Wesley, C.L., White, D., Ye, P.: Curriculum guidelines for undergraduate programs in data science. Annu. Rev. Stat. Appl. **4**(2), 1–16 (2017). URL https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf
413. Vesset, D., Woo, B., Morris, H.D., Villars, R.L., Little, G., Bozman, J.S., Borovick, L., Olofson, C.W., Feldman, S., Conway, S., Eastwood, M., Yezhkova, N.: Worldwide big data technology and services 2012-2015 forecast (2012). IDC
414. Viseu, A., Suchman, L.: Wearable Augmentations: Imaginaries of the Informed Body, pp. 161–184. Berghahn Books, New York (2010)
415. Walker, M.A.: The professionalisation of data science. Int. J. of Data Science **1**(1), 7–16 (2015)
416. Wang, C., Cao, L., Chi, C.: Formalization and verification of group behavior interactions. IEEE Trans. Systems, Man, and Cybernetics: Systems **45**(8), 1109–1124 (2015)
417. WEF: The global competitiveness report 2011-2012: An initiative of the world economic forum (2011)
418. Wei, W.: Copula-based high dimensional dependence modelling. Ph.D. thesis, University of Technology Sydney (2014)
419. Wei Wei Junfu Yin, J.L., Cao, L.: Modeling asymmetry and tail dependence among multiple variables by using partial regular vine. In: SDM2014 (2014)
420. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(1) (2016)
421. Whitehouse: The white house names dr. DJ patil as the first U.S. chief data scientist (2015). URL https://www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist
422. Wikipedia: Bioinformatics. URL https://en.wikipedia.org/wiki/Bioinformatics
423. Wikipedia: Computational trust. URL https://en.wikipedia.org/wiki/Computational_trust
424. Wikipedia: Computing. URL https://en.wikipedia.org/wiki/Computing
425. Wikipedia: Dikw pyramid. URL https://en.wikipedia.org/wiki/DIKW_Pyramid
426. Wikipedia: Genetic linkage. URL https://en.wikipedia.org/wiki/Genetic_linkage
427. Wikipedia: Health care & analytics. URL http://analytics-magazine.org/health-care-a-analytics/
428. Wikipedia: Intelligent transportation system. URL https://en.wikipedia.org/wiki/Intelligent_transportation_system
429. Wikipedia: Social influence. URL https://en.wikipedia.org/wiki/Social_influence
430. Wikipedia: Social network analysis. URL https://en.wikipedia.org/wiki/Social_network_analysis
431. Wikipedia: Statistical relational learning. URL https://en.wikipedia.org/wiki/Statistical_relational_learning
432. Wikipedia: Sustainability. URL https://en.wikipedia.org/wiki/Sustainability
433. Wikipedia: Targeted advertising. URL https://en.wikipedia.org/wiki/Targeted_advertising
434. Wikipedia: Comparison of cluster software (2016). URL https://en.wikipedia.org/wiki/Comparison_of_cluster_software
435. Wikipedia: General data protection regulation (2016). URL https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
436. Wikipedia: Informatics (2016). URL https://en.wikipedia.org/wiki/Informatics
437. Wikipedia: List of reporting software (2016). URL https://en.wikipedia.org/wiki/List_of_reporting_software
438. Wikipedia: National data protection authority (2016). URL https://en.wikipedia.org/wiki/National_data_protection_authority
439. Wikipedia: Sports analytics (2016). URL https://en.wikipedia.org/wiki/Sports_analytics
440. Wikipedia: Accuracy, precision, recall and specificity (2017). URL https://en.wikipedia.org/wiki/Precision_and_recall

441. Wikipedia: Capability maturity model (cmm) (2017). URL https://en.wikipedia.org/wiki/Capability_Maturity_Model
442. Wikipedia: Complexity (2017). URL https://en.wikipedia.org/wiki/Complexity
443. Wikipedia: Data quality (2017). URL https://en.wikipedia.org/wiki/Data_quality
444. Wikipedia: Industrial revolution (2017). URL https://en.wikipedia.org/wiki/Industrial_Revolution
445. Wikipedia: List of statistical packages (2017). URL https://en.wikipedia.org/wiki/List_of_statistical_packages
446. Wikipedia: Second industrial revolution (2017). URL https://en.wikipedia.org/wiki/Second_Industrial_Revolution
447. Wikipedia: Timeline of machine learning. retrieved 21 march 2017 (2017). URL https://en.wikipedia.org/wiki/Timeline_of_machine_learning
448. Wikipedia: Agile software development (2018). URL https://en.wikipedia.org/wiki/Agile_software_development
449. Wikipedia: Industry 4.0 (2018). URL https://en.wikipedia.org/wiki/Industry_4.0
450. Wikipedia: Internet of things (2018). URL https://en.wikipedia.org/wiki/Internet_of_things
451. Wikipedia: Open access (2018). URL https://en.wikipedia.org/wiki/Open_access
452. Wikipedia: Open data (2018). URL https://en.wikipedia.org/wiki/Open_data
453. Wikipedia: Open education (2018). URL https://en.wikipedia.org/wiki/Open_education
454. Wikipedia: Open peer review (2018). URL https://en.wikipedia.org/wiki/Open_peer_review
455. Wikipedia: Open science (2018). URL https://en.wikipedia.org/wiki/Open_science
456. Wikipedia: Open source (2018). URL https://en.wikipedia.org/wiki/Open-source_software
457. Wikipedia: Smart manufacturing (2018). URL https://en.wikipedia.org/wiki/Smart_manufacturing
458. Wikipedia: Waterfall model (2018). URL https://en.wikipedia.org/wiki/Waterfall_model
459. Williamson, J.: Big data analytics is transforming manufacturing (2016). URL http://www.themanufacturer.com/articles/big-data-analytics-is-transforming-manufacturing/
460. WIRED: How europe can seize the starring role in big data (2014). URL www.wired.com/insights/2014/09/europe-big-data/
461. Wladawsky-Berger, I.: Why do we need data science when we've had statistics for centuries? The Wall Street Journal (2014). URL http://blogs.wsj.com/cio/2014/05/02/why-do-we-need-data-science-when-weve-had-statistics-for-centuries/
462. Wolf, G.: The data-driven life. New York Times (2012). URL www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html
463. Woodall P., B.A., Parlikad, A.: Data quality assessment: The hybrid approach. Information & Management **50**(7), 369–382 (2013)
464. Woodall P., O.M., A., B.: A classification of data quality assessment and improvement methods. International Journal of Information Quality **3**(4), 298–321 (2014)
465. Works, B.: Burtch works flash survey (2014). URL http://www.burtchworks.com/category/flash-survey/
466. WTTC: Big data - the impact on travel & tourism (2014). URL https://www.wttc.org/research/other-research/big-data-the-impact-on-travel-tourism/
467. Wu, J.: Statistics = data science? (1997). URL http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf
468. Xie, T., Thummalapenta, S., Lo, D., Liu, C.: Data mining for software engineering. Computer **42**(8) (2009)
469. Yahoo: Yahoo finance (2016). URL www.finance.yahoo.com
470. Yau, N.: Rise of the data scientist (2009). URL http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/
471. Yin, J., Zheng, Z., Cao, L.: Uspan: An efficient algorithm for mining high utility sequential patterns. In: KDD 2012, pp. 660–668 (2012)
472. Yiu, C.: The big data opportunity (2012). URL http://www.policyexchange.org.uk/images/publications/the%20big%20data%20opportunity.pdf
473. Yu, B.: IMS presidential address: Let us own data science. IMS Bulletin Online (2014). 1 Oct 2014

Index

© Springer International Publishing AG, part of Springer Nature 2018
L. Cao, *Data Science Thinking*, Data Analytics,
https://doi.org/10.1007/978-3-319-95092-1