# Chapter 2
# What Is Data Science


## 2.1 Introduction

The art of data science [197] has increasingly attracted interest from a wide range of domains and disciplines. Communities or proposers from diverse backgrounds have often had contrasting aspirations, and have accordingly presented very different views or demonstrated contrasting foci. For example, statisticians may hold the view that data science is the new generation of statistics; people adopting a fresh approach may believe that data science is a consolidation of several interdisciplinary fields, or is a new body of knowledge; industry players may believe that it has implications for providing capabilities and practices for the data profession, or for generating business strategies.

In this chapter, these different views and definitions of data science are discussed. Key concepts including datafication, data quantification, data DNA, and data product are also discussed, as they build the foundation for data science and the world of data. Several definitions of data science are given, followed by discussions about myths and misconceptions about some key concepts and techniques related to data science.

## 2.2 Datafication and Data Quantification

Data is ubiquitous because *datafication* [19] and *data quantification* are ubiquitous. Datafication and data quantification contribute to the recognition of the big data era, and its corresponding challenges and prospects.

*Datafication* refers to how data is created, extracted, acquired or re-produced and rendered into specific data formats, from diverse areas, and through particular channels. Datafication takes the information in every area of business, working, daily life, and entertainment, renders it into data formats, and turns it into data.



L. Cao, *Data Science Thinking*, Data Analytics,
https://doi.org/10.1007/978-3-319-95092-1_2

*Data quantification* (or data quantitation) refers to the act of measuring and counting for the purpose of converting observations to quantities.

In addition to the commonly seen transactions acquired from business and operational information systems, increasingly popular and widespread datafication and data quantification systems and services are significantly contributing to the data deluge and big data realm.

As we have seen and can predict, datafication and data quantification may take place at any time in any place by anybody in any form in any way in a non-traditional manner, to a variable extent and depth, and at fluctuating speed.

- Quantification timing: *anytime quantification*, from working to studying, day-to-day living, relaxing, enjoying entertainment and socializing;
- Quantification places: *anyplace quantification*, from biological systems to physical, behavioral, emotional, cognitive, cyber, environmental, and cultural spaces, and in economic, sociological and political systems and environments;
- Quantification bodies: *anybody quantification*, from selves to others, connected selves, exo-selves [250] and the world, and from individuals to groups, organizations and societies;
- Quantification forms: *anyform quantification*, from observation to drivers, from objective to subjective, from physical to philosophical, from explicit to implicit, and from qualitative to quantitative forms and aspects;
- Quantification ways: *anysource quantification*, such as sources and tools that include information systems, digitalization, sensors, surveillance and tracking systems, the IoT, mobile devices and applications, social services and network platforms, and wearable devices [414] and Quantified Self (QS) devices and services; and
- Quantification speed: *anyspeed quantification*, from static to dynamic, from finite to infinite, and from incremental to exponential generation of data objects, sets, warehouses, lakes and clouds.

Examples of fast developing quantification areas are the health and medical domains. We are datafying both traditional medical and health care data and "omics" data (genomics, proteomics, microbiomics, metabolomics, etc.) and increasingly overwhelming QS-based tracking data [377] on personal, family, group, community, and/or cohort levels.

## 2.3 Data, Information, Knowledge, Intelligence and Wisdom

Before we discuss the definitions of data science, two fundamental questions to answer are:

- *What are data, information, knowledge, intelligence, and wisdom (DIKIW)?* and
- *What are the differences between data, information, knowledge, intelligence, and wisdom?*

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

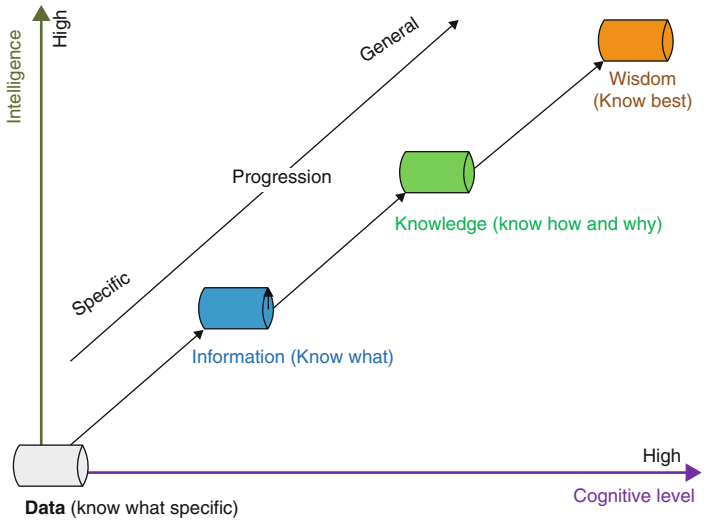


**Fig. 2.1** Data-to-information-to-knowledge-to-intelligence-to-wisdom cognitive progression. Note: X-axis: the increase in cognitive level; Y-axis: the increase in intelligence

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.

Second, *model-independent discovery*, also commonly called *data-driven discovery*, is the correct way to conduct data understanding; that is, to allow the data to tell a story. Models are not applied directly, since a model is built with certain embedded hypotheses and assumptions.

It is not easy to be fully data-driven, since the understanding of complex data characteristics and contexts is often challenging. A more feasible approach is therefore to combine data-driven discovery with model-based learning, which basically assumes that the data fits a certain assumption that can be captured by a model (usually a mathematical or statistical model) while data characteristics have to be deeply explored and then fed to the model. Section 3.5.3.2 discusses this approach.

Lastly, *evidence-based decision-making* is the outcome of data-centric scientific exploration. Here *evidence* constitutes the new hypotheses, insights, indicators, and findings hidden in data that were previously invisible to data modelers and decision makers.

Evidence is not simply the modeling of outputs or results; rather, it is the deep disclosure of the intrinsic nature, characteristics, complexities, principles, and stories inbuilt in the data and physical world. In essence, data-driven evidence captures the working mechanism, intrinsic nature, and solutions of the underlying problem.

The data-centric view is also largely derived from a large number of conceptual arguments; for example, that data-driven science is mainly interpreted in terms of the reuse of open data [299, 312], that data science comprises the numbers of our lives [290], or that data science enables the creation of data products [278, 279].

## 2.6 Definitions of Data Science

The various data science views discussed in Sect. 2.5 form the foundation for our discussion on what data science is. Below, we present several definitions of data science from different perspectives, building on the observations and insights we have gained from a comprehensive review of data science [63–66] and relevant experience in advanced analytics over the past decades, as well as inspirations from the relevant discussions about the concepts and scopes of data science (e.g., in [12, 19, 93, 97, 112, 128–130, 197, 205, 215, 231, 238, 279, 283, 285–287, 316, 327, 332, 364, 371, 461, 467]).

### *2.6.1 High-Level Data Science Definition*

Taking the data-centric view outlined in Sect. 2.5.3, a high-level definition of data science can be adopted that accords with the usual type of umbrella definition found in any scientific field, as given below:

**Definition 2.2 (Data Science[1])** Data science is the science of data, or data science is the study of data.

However, this high-level view does not provide a concrete explanation of what makes data science a new science, nor why or how. Therefore, in the following sections, more concrete and specific data science definitions are given.

### *2.6.2 Trans-Disciplinary Data Science Definition*

As the multi- and trans-disciplinary views of data science indicate, *data science* is a new disciplinary field in which to study data and its domain with data science thinking (more discussion about data science thinking can be found in Chap. 3.).

**Definition 2.3 (Data Science[2])** Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, such as statistics, informatics, computing, communication, management and sociology, to study data and its domain employing data science thinking.

Since a variety of multidisciplinary views exist, as discussed in Sect. 2.5.2, different people may have contrasting opinions of which disciplines constitute data science. In general, data science integrates *traditionally data-oriented disciplines* such as statistics, informatics, and computing with *traditionally data-independent fields* such as communication, management, and sociology.

As an example (illustrated in Fig. 2.2), a *discipline-based data science formula* is given below:

$$\begin{aligned} &data\ science \\ &\quad \stackrel{\text{def}}{=} \{statistics \cap informatics \cap computing \cap communication \\ &\qquad \cap sociology \cap management \mid data \cap domain \cap thinking\} \end{aligned} \tag{2.1}$$

where "|" means "conditional on."

*Data science thinking* is required to understand what data science is and to generate data products [64]. This thinking reflects data-driven methodologies and process, and the transformation from data analytics to knowledge generation and wisdom production.

In the data-to-information-to-knowledge-to-intelligence-to-wisdom progression (see Sect. 2.3), different disciplines play different roles.

In Chap. 6, we briefly discuss the roles played by relevant disciplines and the inter-disciplinary, cross-disciplinary and trans-disciplinary interactions between data science and the respective disciplines.

The definition of Data Science[2] also includes the new opportunities and development that exist through migrating or match-making existing disciplinary development to a data science-oriented stage. This highlights the potential for new data

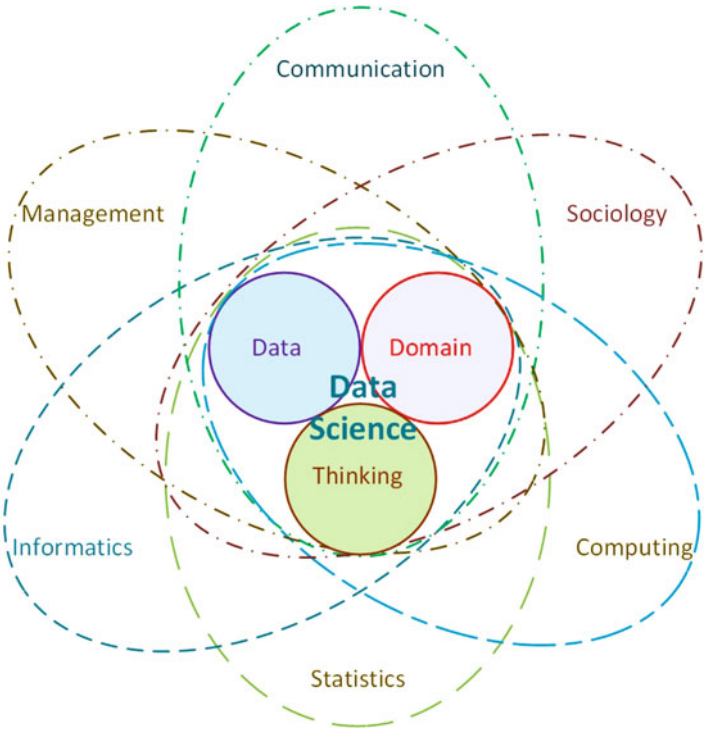


**Fig. 2.2** Trans-disciplinary data science

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".

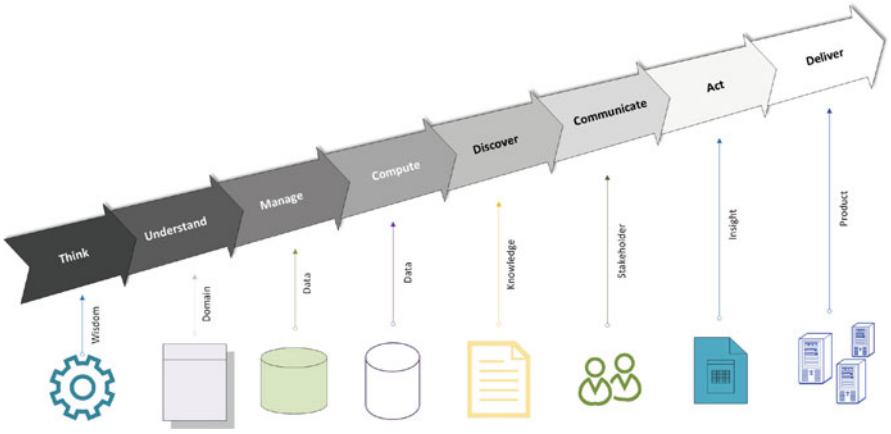


**Fig. 2.3** Process-driven data science

As shown in Fig. 2.3, data science involves many important activities and corresponding resources and facilities related to DIKIW. Accordingly, a *process-based data science formula* is given below:

$$data\ science \stackrel{\text{def}}{=} \{think + understand + manage + compute + discover + communicate + act + deliver | DIKIW\} \quad (2.2)$$

These components have specific meanings in the data science process, and these are discussed in the following subsections.

#### 2.6.3.1 Thinking with Wisdom

*Thinking with wisdom* reflects the driving force, goal/objective and outcomes of data science that will transform data to wisdom. Data is the input, while wisdom reflects the data science outcomes.

Thinking with wisdom also reflects the value of data science, which is to produce wisdom for smarter, evidence-based, more informative decision-making. Thinking with wisdom reflects the challenge of data science, how much wisdom we can obtain from data-driven discovery, and what wisdom we can obtain through data science.

Thinking with wisdom is an output-driven thinking and purpose-based methodology. It differentiates data science from existing sciences such as statistics, computing and informatics in the sense that data science expects to produce deep insights that are not visible, and are evidence-based and valuable.

Initial wisdom comes from experience, expert knowledge, and preliminary hypothesis about the input data, problem, goal and path for problem-solving. These elements may be incorrect, and will also be fully domain-dependent. They will

be verified, refined, updated or renewed with confirmed insights and intelligence following the completion of the data science process.

#### 2.6.3.2 Understanding the Domain

Although there are general scientific and technological issues and systems, data science is often domain-dependent when the data science process is focused. For example, network science consists of data science that is customized to address networking complexities, characteristics, and objectives.

Domain-specific data science is necessary because data is usually domain-specific. This is similar to X-analytics, in which analytics are developed for specific domains; the same is true for X-informatics, although general analytics and informatics problems and theories are also essential.

In creating domain-specific data science, such as behavioral data science, domain complexities and characteristics need to be scrutinized and articulated. Specifying and generalizing domain-independent complexities and characteristics lead to general data science; specifying and generalizing domain-dependent complexities and characteristics lead to domain-specific data science.

As discussed in Chap. 6, domain-specific data science needs to consider many domain factors, domain complexities, and domain intelligence. In Sect. 2.6.2, the roles of and relations between data, domain and thinking in relation to what data science is are further discussed.

#### 2.6.3.3 Managing Data

*Managing data* is a core component of data science. It involves the acquisition, management, storage, search, indexing and recovery of data.

The pre-processing of data is critical in managing data, and is related to data quality enhancement and the management of social issues.

Data quality enhancement involves the handling of data quality problems, such as missing values, errors, and fake data, and the processing of data complexities such as long-tail distributions and sparsity, bias and odd distributions, and non-IID characteristics.

Social and ethical issues are increasingly important in data science. Typical issues that need to be managed include privacy processing and protection, the security of data and systems, and the trust of data and data-based products. Detailed discussion about data quality, data quality issues, and data quality assurance is available in Sect. 4.7. Data social issues and data ethics, and the assurance of both, are discussed in Sect. 4.8.

In addition to data management, the role of management science in data science is also discussed in Sect. 6.8.

#### 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.

Data products are highly variable, depending on their purpose, the specific procedures of the data science process, the data science personnel who produce the deliverables, overall requirements and expectations, and so on.

Findings from data mining and machine learning are often presented in a technical manner and as half-way products, which need to be converted to final products, such as by incorporating business impact and being presented in business rules that are easily acceptable for business operations and decision support.

In business, data products extend far beyond data analytical models and algorithms. A data product may be presented as a mobile application, a service website, a social media system, a dashboard, an infrastructure-oriented system, a programming language, or even an analytical report.

See more discussion on the topic of data products in Sects. 2.8 and 1.3.1.

#### 2.6.3.8 Acting on Insights

Insights in data science, i.e., *data insights*, are presented in two forms: analytical insight and deliverable insight.

The most valuable and difficult thing in data science is to extract insights from data and reflect such insights in data science deliverables, namely *deliverable insights*. *Deliverable insights* refer to the deep understanding and appropriate representation of intrinsic, fundamental, genuine and complete mechanisms, dynamics, principles, and driving forces in a business problem, and are reflected in the acquired data.

Deep analytics converts data into insights for decision-making. *Deep insights* discover original working mechanisms mapped from the physical world to the data world.

The second type of insight is the analytical insight, which is often overlooked in data science. Converting data to knowledge and wisdom requires deep insights, which drive and enable the deep understanding of data characteristics and complexities.

*Analytical insights* provide the appropriate perspectives to effectively capture the intrinsic and fundamental mechanisms, dynamics, principles and driving factors reflected in the data world. Naturally, analytical insights have to be extracted by X-analytics (more discussion in Sect. 7.6) by incorporating and utilizing X-complexities (see Sect. 4.2) and X-intelligence (see Sect. 4.3).

## 2.7 Open Model, Open Data and Open Science

The power of data continues to significantly transform the philosophy, methodology, mechanism and process of traditional data and science development, management, and evolution. The revolution in science, technology and economy is driven by an open model that drives many open movements and activities, in particular, open data

and open science. These are built on the spirit of openness. This section discusses these critical enablers of the fourth revolution—data-driven science, technology, and economy.

### 2.7.1 Open Model

*Openness* is a critical system feature of open complex systems and their evolution, in which there are interactions and exchanges of information, energy, or materials between systems and outside environments [62, 333]. Systems with openness are open systems. Openness has shown that it plays a critical role in driving the faster development of an open system compared to a closed system without openness.

A key feature differentiating the data science era from the pre-data science era is the overwhelming adoption and acceptance of an open model rather than a closed one. The *open model* is enabled by adopting the openness principle, and advocates and enables public, free or shared, democratic, distributed, and collaborative modes in every aspect of science, technology, economy, society, and living. We outline the important features of each mode below.

- Public: is accessible to a wide community and audience;
- Free: is freely available without a financial paywall or for-profit publishing;
- Shared: is accessible with certain restrictions;
- Democratic: is contributed by and contributes to the public; and
- Collaborative: is jointly contributed by and jointly contributes to its audience.

Typical examples of open model-driven technological, economic, societal and lifestyle entities include the innovation of social media such as Facebook and LinkedIn, the migration of mobile to smart phone-embedded applications, and industrial transformations such as the migration of physical shop-based commerce to online businesses like Amazon and Taobao.

The openness principle and open model have been rapidly adopted and developed since Internet-based and smart phone applications have become widespread in everyday business. Many open model-based movements (open movements) and activities (open activities) have emerged, such as open source, open data, open access, open government, open education, and open science, which has fundamentally transformed their existing principles, mechanisms, operations, stakeholder relationships, and financial and accounting models. These open systems and mechanisms have been globalized and are represented by worldwide Internet-based movements, services, and practices, which interact with each other and their corresponding classic systems and mechanisms, driving the rapid evolution of our science, technology, economy, and society and living towards a more open and universal way of conducting business and communicating.

### 2.7.2 *Open Data*

Open data [452] is a fundamental principle and mechanism in the open model. It encourages and enables the free availability of data, particularly through the Internet, social media and mobile applications-based technologies, infrastructure, services, and operations. *Open data* offers the public access to, usage, dissemination and republishing of data without no copyright or financial paywall demands other than acknowledgement. Open data mechanisms and activities thus remove the control typically applied in classic closed systems, which include restrictions on access, intellectual property (including copyright, patents), and for-profit business models (e.g., licences).

There is an expanding variety of data that can be made open. Typical open data consist of scientific data (from every field of science), common and professional knowledge (from the public domain, literature and disciplinary research), policies (from institutions, government and professional bodies), and the public (about their activities and sentiment).

Open data and data sharing programs have been announced in many countries and domains. Major developments are the open government data initiatives supported by government organizations, and the corresponding policy support; for example, the US Government open data site [403], the UK open data project [396, 397], the Australian Government open government data site [15, 16] and Data Matching program [14], and the European Union Open Data Portal [155] and data sharing projects [212].

In addition, many Open Access schemes are increasingly being accepted by academic journals and funding bodies, and are included in institutional evaluations.

Efforts have also been made in diverse societies to create shareable data repositories, especially for science and research. Examples of open repositories are the global climate data [389], the global terrorism database [199], the Yahoo Finance data [469], the Gene Expression Omnibus [179], mobile data [194], the UCI repositories for machine learning [391], the Linguistic Data Consortium data for Natural Language Processing [270], the TREC data for text retrieval [309], Kaggle competition data [241], and the VAST challenge [411] for visual analytics, to name a few.

However, it is understandable that not all data is shareable, and exceptions need to be made for private data and other sensitive data. Open data also does not mean that the data can be used for unlawful purposes. The amount of non-open data is reducing, but issues related to the ethics, privacy and misuse of open data are emerging that could prove to be very serious, such as the 2018 Facebook scandal in which Cambridge Analytica improperly utilized Facebook user data in their commercial activities [24]. Relevant national and international laws, policies, agreements, and norms are not yet sufficiently well developed to regulate and support healthy open data activities and to protect the data creator/owner's information and rights from misuse. When powerful data is misused, negative outcomes and impact may result, so an awareness of the need to prevent such

outcomes should be foremost in promoting open data and open activities. Urgent and systematic research, regulation and education on the risks and mitigation of open data and related open activities are critical from a legal, ethical, political, economic, and behavioral perspective.

### *2.7.3 Open Science*

The scientific community has been one of the very first public domains to welcome, advocate and promote the openness principle and open model-driven activities in science, research, innovation, data management, dissemination of results, and intellectual property management. This has driven the formation of open science [455]. Open science takes the openness principle and implements the open model in all these areas, and in related activities (e.g., knowledge sharing, reviewing and evaluation, publishing and republishing, education, operations, and management).

Typically, open science consists of the following major open mechanisms, methods and activities: open research and innovation, open science data and open access, open source, open review and evaluation, and open education and training, as shown in Fig. 2.4.

*Open research and innovation* are the core activities in open science. They advocate and drive the undertaking of scientific research and innovation activities in an open fashion. Open research and innovation are more transparent, collaborative, and distributed than closed research which is undertaken within a more independent, confidential and private context. Open research highlights the spirit and value of conducting collaborative and shareable scientific activities, which drives the scientific activities of problem identification and statement, the formulation and


**Fig. 2.4** Open science

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

the assessment and refinement of scientific outputs and impact. Open science is significantly driving the development of data science as a new science.

## 2.8 Data Products

The outputs of data science are *data products* [278, 279]. Data products can be described as follows.

**Definition 2.5 (Data Products)** *Data products* are deliverables and outputs from data, or driven by data. Data products can be presented as discoveries, predictions, services, recommendations, decision-making insights, thinking, models, algorithms, modes, paradigms, tools, systems, or applications. The ultimate data products of value are knowledge, intelligence and wisdom.

The above definition of data products goes beyond technical product-based types and forms in the business and economic domain, such as social network platforms like Facebook, and recommender systems like Netflix. Producing data-based outputs in terms of products and business enables the generation of a new economy: *data economy*. Typically, the current stage of data economy is featured by social data business, mobile data business, online data business, and messaging-based data business.

Various data products enable us to explore new data-driven or data-enabled personalized, organizational, educational, ethical, societal, cultural, economic, political, cyber-physical forms, modes, paradigms, innovations, directions and ecosystems, or even thinking, strategies and policies. For example, there is a good possibility that large scale data will enable and enhance the transfer of subjective autonomy to objective autonomy, beneficence and justice in the social sciences [158], and will enable the discovery of indicators like Google Flu [269] which may not be readily predicted by domain-driven hypothesis and professionals.

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

## 2.9 Myths and Misconceptions

Data science is still at an early stage, and it is therefore understandable to see different and sometimes contradictory views appearing in various quarters. However, it is essential to share and discuss the many myths, memes [129], and

misconceptions about data science compared to the reality [234], and to ensure the healthy development of the field. This section draws on observations about the relevant communities, as well as the experiences and lessons learned in conducting data science and analytics research, education, and services, to list the relevant myths and misconceptions for discussion.

At the same time, it is important to debate the nature of data science, clarify the fundamental concepts and myths, and demonstrate the intrinsic characteristics and opportunities that data science has to offer.

Hence, this section lists various misunderstandings and myths about data science and also clarifies the reality. The common misconceptions are summarized in terms of the concepts of data science, data volume, infrastructure, analytics, capabilities and roles. Discussion and clarification are provided to present the actual status, intrinsic factors, characteristics, and features of data science, as well as the challenges and opportunities in data research, disciplinary development, and innovation.

### *2.9.1 Possible Negative Effects in Conducting Data Science*

Big data and data science are fundamental drivers in the new generation of science, technologies, economy, and society. However, this does not mean that all applications and data science case studies will naturally have a positive effect, as might be expected.

While often overlooked, the improper use of big data and data science may result in negative effects. These may be caused such behaviors as

- the violation of assumptions made in applied theories, models, tools, and results;
- the improper alignment of theories and methods with their applicability (corresponding context and conditions);
- the violation of constraints (applicability) on data characteristics and complexities to fit the applied theories, models, and results;
- the quality issues associated with data (e.g., bias, see more in Sect. 4.7);
- the applicability and potential of data;
- the quality issues associated with features;
- the quality issues associated with theories and methods;
- the applicability and potential of theories and methods;
- the applicability and potential of evaluation systems;
- the availability of theoretical foundations and guarantee of the methods and results;
- the social issues and ethics associated with data science (see discussion in Sect. 4.8);
- the applicability of the resultant output and findings;
- the significance, transparency and learnability of underlying problems and their complexities;

- the effectiveness of computing infrastructure;
- the efficiency of computing infrastructure;

Any of the above scenarios could easily result in biased, misleading, inappropriate or incorrect applications and consequences. The use of biased findings, the overuse, underuse and misuse of modeling assumptions, treating unverifiable outcomes objectively, or enlarging the values of data-driven discovery, may cause defects and problems.

The possible negative effects of the improper application of data science may become particularly obvious and hard to address. One example is in the analysis of complex social problems. There are many sophisticated complexities to be considered in social data science, such as cultural factors, socio-economic and political aspects, in addition to data characteristics and complexities. Their working mechanisms, behaviors, dynamics, and evolution may be much more sophisticated and demonstrate characteristics different from those in business transactions collected from business operations and production.

### *2.9.2 Conceptual Misconceptions*

Data science has typically been defined in terms of specific disciplinary foundations, principles, goals, inputs, algorithms and models, processes, tools, outputs, applications, and professions. Often, however, a fragmented statement may cause debate as a result of the failure to see the whole picture. In this section, we discuss some of the arguments and observations collected from the literature.

- Data science is statistics [44, 128]; "why do we need data science when we've had statistics for centuries" [461]? How does data science really differ from statistics [129]? (Comments: Data science provides systematic, holistic and multi-disciplinary solutions for learning explicit and implicit insights and intelligence from complex and large-scale data, and generates evidence or indicators from data by undertaking diagnostic, descriptive, predictive and/or prescriptive analytics, in addition to supporting other tasks on data such as computing and management.)
- Why do we need data science when information science and data engineering have been explored for many years? (Comments: Consider the issues faced in related areas by the enormity of the task and the parallel example of enabling a blind person to recognize an animal as large as an elephant (see more about blind knowledge and the parallel example in Sect. 4.4.2). Information science and data engineering alone cannot achieve this. Other aspects may be learned from the discussion about greater or fewer statistics; more in [87].)
- I have been doing data analysis for dozens of years; data science has nothing new to offer me. (Comments: Classic data analysis and technologies focus mostly on explicit observation analysis and hypothesis testing on small and simpler data.)

- Is data science old wine in a new bottle? What are the new grand challenges foregrounded by data science? (Comments: The analysis of the gaps between existing developments and the potential of data science (see Fig. 5.1) shows that many opportunities can be found to fill the theoretical gaps when data complexities extend significantly beyond the level that can be handled by the state-of-the-art theories and systems, e.g., classic statistical and analytical theories and systems were not designed to handle the non-IIDness [60] in complex real-life systems.)
- Data science mixes statistics, data engineering and computing, and does not contribute to breakthrough research. (Comments: Data science attracts attention because of the significant complexities in handling complex real-world data, applications and problems that cannot be addressed well by existing statistics, data engineering and computing theories and systems. This drives significant innovation and produces unique opportunities for generating breakthrough theories.)
- Data science is also referred to as data analytics and big data [6]. (Comments: This confuses the main objectives, features, and scope of the three concepts and areas. Data science needs to be clearly distinguished from both data analytics and big data.)
- Other definitions wrongly ascribed to data science are that it is big data discovery [121], prediction [126], or the combination of principle and process with technique [332].

It is also worth noting that the terms *big data*, *data science* and *advanced analytics* are often overused or improperly used by many communities and for various purposes, particularly because of the influence of media hype and buzz. Most Google searches on these keywords return results that are irrelevant to their intrinsic semantics and scope, or simply repeat familiar arguments about the needs of data science and existing phenomena. In many such findings [7, 45, 83, 89, 93, 112, 127, 159, 186, 200, 203, 208, 215, 244, 256, 279, 282, 290, 297, 316, 320, 331, 359, 371, 372], big data is described as being simple, data science is said to have nothing to do with the science of data, and advanced analytics is described as being the same as classic data analysis and information processing. There is a lack of deep thinking and exploration of why, what, and how these new terms should be defined, developed, and applied.

In [234], six myths were discussed:

- Size is all that matters;
- The central challenge of big data is that of devising new computing architectures and algorithms;
- Analytics is the central problem of big data;
- Data reuse is low hanging fruit;
- Data science is the same as big data; and
- Big data is all hype.

This illustrates the constituents of the ecosystem, but also shows the divided views within the communities.

These observations illustrate that data science is still young. They also justify the urgent need to develop sound terminology, standards, a code of conduct, statements and definitions, theoretical frameworks, and better practices that will exemplify typical data science professional practices and profiles.

### *2.9.3 Data Volume Misconceptions*

There are various misconceptions surrounding data volume. For example,

- What makes data "big"? (Comments: It is usually not the volume but the complexities, as discussed in [62, 64], and large values that make data big.)
- Why is the bigness of data important? (Comments: The bigness of data âĂŞ which refers to data science complexities—heralds new opportunities for theoretical, technological, practical, economic and other development or revolution.)
- Big data refers to massive volumes of data. (Comments: Here, "big" refers mainly to significant data complexities. From the volume perspective, a data set is big when the size of the data itself becomes a quintessential part of the problem.)
- Data science is big data analytics. (Comments: Data science is a comprehensive field centered on manipulating data complexities and extracting intelligence, in which data can be big or small and analytics is a core component and task.)
- I do not have big data so I cannot do big data research. (Comments: Most researchers and practitioners do not have sizeable amounts of data and do not have access to big infrastructure either. However, significant research opportunities still exist to create fundamentally new theories and tools to address respective X-complexities and X-intelligence.)
- The data I can find is small and too simple to be explored. (Comments: While scale is a critical issue in data science, small data, which is widely available, may still incorporate interesting data complexities that have not been well addressed. Often, we see experimental data, which is usually small, neat and clean. Observational data from real business is live, complex, large and frequently messy.)
- I am collecting data from all sources in order to conduct big data analytics. (Comments: Only relevant data is required to achieve a specific analytical goal.)
- It is better to have too much data than too little. (Comments: While more data generally tends to present more opportunities, the amount needs to be relevant to the data required and the data manipulation goals. Whether bigger is better depends on many aspects.)

### 2.9.4 Data Infrastructure Misconceptions

Below, we list two misconceptions related to data infrastructure.

- I do not have big infrastructure, so I cannot do big data research. (Comments: While big infrastructure is useful or necessary for some big data tasks, theoretical research on significant challenges may not require big infrastructure.)
- My organization will purchase a high performance computer to support big data analytics (Comments: Many big data analytics tasks can be successfully undertaken without a high performance computer. It is also essential to differentiate between distributed/parallel computing and high performance computing.)

### 2.9.5 Analytics Misconceptions

There are many misconceptions relating to analytics. We list some here for discussion.

- Thinking data-analytically is crucial for data science. (Comments: Data-analytic thinking is not only important for a specific problem-solving, but is essential for obtaining a systematic solution and for a data-rich organization. Converting an organization to think data analytically gives a critical competitive advantage in the data era.)
- The task of an analyst is mainly to develop common task frameworks and conduct inference [42] from the particular to the general. (Comments: Analytics in the real world is often specific. Focusing on certain common task frameworks may trigger incomplete or even misleading outcomes. As discussed in Sect. 10.3.1, an analyst may take other roles, e.g., predictive modeling is typically problem-specific.)
- I only trust the quality of models built in commercial analytical tools. (Comments: Such tools may produce misleading or even incorrect outcomes if the assumption of their theoretical foundation does not fit the data, e.g., if they only suit imbalanced data, normal distribution-based data, or IID data.)
- Most published models and algorithms and their experimental outcomes are not repeatable. (Comments: Such works seem to be more hand-crafted rather than manufactured. Repeatability, reproducibility, open data and data sharing are critical to the healthy development of the field.)
- I want to do big data analytics, can you tell me which algorithms and program language I should learn? (Comments: Public survey outcomes give responses to such questions; see examples in [63]. Which algorithms, language and platform should be chosen also depends on organizational maturity and needs. For long-term purposes, big data analytics is about building competencies rather than specific functions).

- My organization's data is private, thus you cannot be involved in our analytics. (Comments: Private data can still be explored by external parties by implementing proper privacy protection and setting up appropriate policies for onsite exploration.)
- Let me (an analyst) show you (business people) some of my findings which are statistically significant. (Comments: As domain-driven data mining [77] shows, many outcomes are often statistically significant but are not actionable. An evaluation of those findings needs to be conducted to discover what business impact [79] might be generated if the findings they generate are operationalized.)
- Strange, why can I not understand and interpret the outcomes? (Comments: This may be because the problem has been misstated, the model may be invalid for the data, or the data used is not relevant or correct.)
- Your outcomes are too empirical without theoretical proof and foundation. (Comments: While it would be ideal if questions about the outcomes could be addressed from theoretical, optimization and evaluation perspectives, real-life complex data analytics are often more exploratory, and it may initially be difficult to optimize empirical performance.)
- My analysis shows that what you delivered is not the best for our organization. (Comments: It may be challenging to claim "the best" when a variety of models, workflows and data features are used in analytics. It is not unusual for analysts to obtain different or contradictory outcomes on the same data as a result of the application of different theories, settings and models. It may turn out to be very challenging to find a solid model that perfectly and stably fits the invisible aspect of data characteristics. It is important to appropriately check the relevance and validity of the data, models, frameworks and workflows available and used. Doing the right thing at the right time for the right purpose is a very difficult task when attempting to understand complex real-life data and problems.)
- Can your model address all of my business problems? (Comments: Different models are often are required to address diverse business problems, as a single model cannot handle a problem sufficiently well.)
- This model is very advanced with solid theoretical foundation, let us try it in your business. (Comments: While having solid scientific understanding of a model is important, it is data-driven discovery that may better capture the actual data characteristics in real-life problem solving. A model may be improperly used without a deep understanding of model and data suitability. Combining data driven approaches with model driven approaches may be more practical.)
- My analytical reports consist of lots of figures and tables that summarize the data mining outcomes, but my boss seems not so interested in them. (Comments: Analytics is not just about producing meaningful analytical outcomes and reports; rather, it concerns insights, recommendations and communication with upper management for decision-making and action.)
- It is better to have advanced models rather than simple ones. (Comments: Generally, simple is better. The key to deploying a model is to fit the model to the data while following the same assumption adopted by the model.)

- We just tuned the models last month, but again they do not work well. (Comments: Monitoring a model's performance by watching the dynamics and significant changes that may take place in the data and business is critical. Real-time analytics requires adaptive and automated re-learning and adjustment.)
- I designed the model, so I trust the outcomes. (Comments: The reproducibility of model outcomes relies on many factors. A model that is properly constructed may fall short in other aspects such as data leakage, overfitting, insufficient data cleaning, or poor understanding of data characteristics and business. Similarly, a lack of communication with the business may cause serious problems in the quality of the outcome.)
- Data science and analytics projects are just other kinds of IT projects. (Comments: While data projects share many similar aspects with mainstream IT projects, certain distinctive features in data, the manipulation process, delivery, and especially the exploratory nature of data science and analytics projects require different strategies, procedures and treatments. Data science projects are more exploratory, ad hoc, decision-oriented and intelligence-driven.)

### *2.9.6 Misconceptions About Capabilities and Roles*

There are various misconceptions about data science capabilities and roles. For example:

- I am a data scientist. (Comments: Lately, it seems that everyone has suddenly become a data scientist. Most data scientists simply conduct normal data engineering and descriptive analytics. Do not expect omnipotence from data scientists.)
- "A human investigative journalist can look at the facts, identify what's wrong with the situation, uncover the truth, and write a story that places the facts in context. A computer can't." [256] (Comments: The success of AlphaGo and AlphaGo Zero [189] may illustrate the potential that a data science-enabled computer has to undertake a large proportion of the job a journalist does.)
- My organization wants to do big data analytics, can you recommend some of your PhD graduates to us? (Comments: While data science and advanced analytics tasks usually benefit from the input of PhDs, an organization requires different roles and competencies according to the maturity level of the analytics and the organization.)
- Our data science team consists of a group of data scientists. (Comments: An effective data science team may consist of statisticians, programmers, physicists, artists, social scientists, decision-makers, or even entrepreneurs.)
- A data scientist is a statistical programmer. (Comments: In addition to the core skills of coding and statistics, a data scientist needs to handle many other matters; see the discussion in [63].)

### 2.9.7 Other Matters

Some additional matters require careful consideration in relation to conducting data science and analytics. We list a few common remarks, with our comments in parentheses.

- Can big data address or even solve big questions, such as global issues, global warming, climate change, terrorism, financial crises? (Comments: Mixed sources of linked data and data matching enable the comprehensive analysis of big questions and global issues, for example, linking airfare booking, accommodation booking, transport information, and cultural background.)
- Can data science transform existing disciplines such as social science? (Comments: When data science meets social science, the synergy between the two has a good chance of promoting the upgrade and transformation of social science, in particular sociology, by providing new thinking, new methods, and new approaches, such as reductionism, data-driven discovery, the scaling-up of social experiments, deep analytics of social problems, and crowdsourcing-based problem-solving.)
- Garbage in, garbage out. (Comments: The quality of data determines the quality of output.)
- More complex data, a more advanced model, and better outcomes. (Comments: Good data does not necessarily lead to good outcomes; a good model also does not guarantee good outcomes.)
- More general models, better applicability. (Comments: General models may lead to weaker outcomes on a specific problem. It is not reasonable or practical to expect a single tool for all tasks.)
- More frequent patterns, more interesting. (Comments: It has been shown that frequent patterns mined by existing theories are generally not useful and actionable.)
- We're interested in outcomes, not theories. (Comments: Actionable outcomes may need to satisfy both technical and business significance [77].)
- The goal of analytics is to support decision-making actions, not just to present outcomes about data understanding and analytical results. (Comments: This addresses the need for actionable knowledge delivery [54] to recommend actions from data analytics for decision support.)
- Whatever you do, you can at least get some values. (Comments: This is true, but it may be risky or misleading. Informed data manipulation and analytics requires a foundation for interpreting why the outcomes look the way they do.)
- Many end users are investing in big data infrastructure without project management. (Comments: Do not rush into data infrastructure investment without a solid strategic plan of your data science initiatives, which requires the identification of business needs and requirements, the definition of reasonable objectives, the specification of timelines, and the allocation of resources.)
- Pushing data science forward without suitable talent. (Comments: On one hand, you should not simply wait for the right candidate to come along, but should

actively plan and specify the skills needed for your organization's initiatives and assemble a team according to the skill-sets required. On the other hand, getting the right people on board is critical, as data science is essentially about intelligence and talent.)
- No culture for converting data science insights into actionable outcomes. (Comments: This may be common in business intelligence and technically focused teams. Fostering a data science-friendly culture requires a top-down approach driven by business needs, making data-driven decisions that enable data science specialists and project managers to be part of the business process, and conducting change management.)
- Correct evaluation of outcomes. (Comments: This goes far beyond such technical metrics as Area Under the ROC Curve and Normalized Mutual Information. Business performance after the adoption of recommended outcomes needs to be evaluated [54]. For example, recent works on high utility analysis [471] and high impact behavior analysis [79] study how business performance can be taken into account in data modeling and evaluation. Solutions that lack business viability are not actionable.)
- Apply a model in a consistent way. (Comments: It is essential to understand the hypothesis behind a model and to apply a model consistent with its hypothesis.)
- Overthinking and overusing models. (Comments: All models and methods are specific to certain hypotheses and scenarios. No models are universal and sufficiently "advanced" to suit everything. Do not assume that if the data is tortured long enough, it will confess to anything.)
- Know nothing about the data before applying a model. (Comments: Data understanding is a must-do step before a model is applied.)
- Analyze data for the sake of analysis only. (Comments: This involves the common bad practice of overusing analytics.)
- What makes an insight (knowledge) actionable? (Comments: This is dependent on not only the statistical and practical values of the insight, but also its predictive power and business impact.)
- Do not assume the data you are given is perfect. (Comments: Data quality forms the basis for obtaining good models, outcomes and decisions. Poor quality data, the same as poor quality models, can lead to misleading or damaging decisions. Real-life data often contains imperfect features such as incompleteness, uncertainty, bias, rareness, imbalance and non-IIDness.)

The above is only a partial list of the misconceptions and misunderstandings that can be witnessed and heard in the current data science and analytics community.

It is not surprising to hear a range of diverse arguments, ideas, and beliefs about any new area. Data science is a highly non-traditional, interdisciplinary, cross-domain, transformative, fast-growing, and pragmatic scientific field. It is in fact a very positive sign to see the increased debate and emerging understanding of the intrinsic nature of data science as a new science, and data economy as the new generation of economy. Importantly, significant and deterministic efforts are required to substantially deepen and widen our understanding and repetitively reflect

on our understanding of data science, data innovation, and data economy. In doing so, we will hopefully capture the true value of data science and guide its evolution to a substantial new stage.

## 2.10 Summary

The answers to the question "What is data science?" are varied, and sometime confusing and conflicting. Exploring data science understanding from a range of aspects and perspectives provides more opportunities to explore and observe the nature of data science.

This chapter first explores how data is quantified (namely by datafication and data quantification), the relationships between five key concepts: data, information, knowledge, intelligence and wisdom, introduces the concept of data DNA as the key datalogical "molecule" of data, and explains the outputs of data science: data products.

This knowledge lays the foundation for exploring the different meanings of data science, the multiple views on data science, and the various definitions of data science. Lastly, this chapter lists and comments on many myths and misconceptions appearing in literature.

The next important and challenging question is "What makes data science a new science?" Answering this question requires an in-depth understanding of the unique but fundamental abstraction in data science, otherwise known as "data science thinking". Accordingly, Chap. 3 explains the concept of data science thinking and discusses its implications.