Chapter 7
Data Science Techniques

7.1 Introduction

In the age of analytics, data analytics and learning form a comprehensive spectrum and evolutionary map that cover

- the whole life cycle of the data from the past to the present, and into the future,
- the analytics and learning from the perspective of known and reactive understanding to unknown and proactive early prediction and intervention, and
- the journey from data exploration to the delivery of actionable [77] insights through descriptive-to-predictive-prescriptive analytics.

In this new era of data science, critical questions to be answered are

- what are the major data characteristics and complexities?
- what is to be analyzed?
- what constitutes the analytics spectrum for understanding data? and
- what form does the paradigm shift of analytics take?

The focus in this chapter, in addition to addressing these questions, is on providing a review and overview of the key tasks, approaches, and lessons associated with various stages, forms and methods of analytics and learning. Highlighted are the paradigm shift and transformation

- from data to insight, and decision-making;
- from explicit to implicit analytics;
- from descriptive to predictive and prescriptive analytics; and
- from shallow learning to deep learning and deep analytics.

Before introducing the above, a high level overview of data science techniques is given. Lastly, general discussion on the marriage of analytics with specific domains, that is, forming X-analytics, is also provided. The chapter concludes with a discussion of the relevant analytics and learning techniques, approaches and tasks.

L. Cao, *Data Science Thinking*, Data Analytics,
https://doi.org/10.1007/978-3-319-95092-1_7

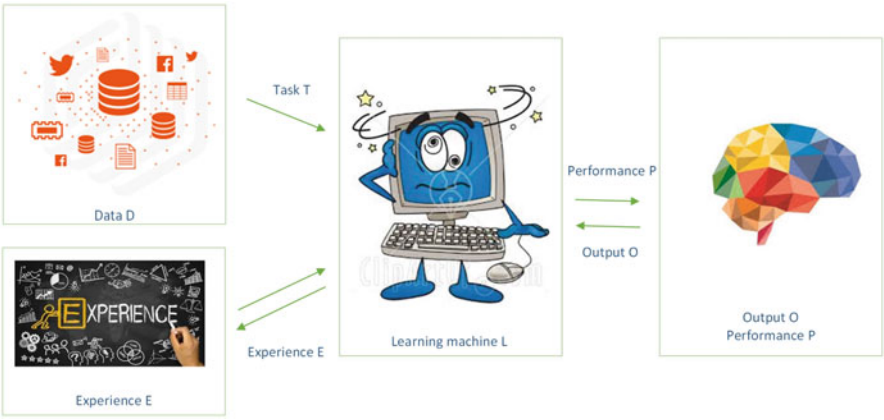

Fig. 7.1 The definition of general machine learning tasks. Note: the diagrams embedded in this figure are from Google search

7.2 The Problem of Analytics and Learning

Data analytics and machine learning are subfields of computer science which play a fundamental role in the innovation and development of modern artificial intelligence and machine intelligence systems.

As shown in Fig. 7.1, a machine learning method is often expressed as a computer program to "learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [295].

A machine learning activity generates a suitable *learning machine* L which obtains the best performance P in undertaking task T on data D, based on experience E. An objective function F is defined to achieve the optimum performance P of solving the learning task T.

7.3 The Conceptual Map of Data Science Techniques

Since data analysis and machine learning were proposed several decades ago, analytics and learning have experienced a significant transformation in their development in terms of target problems, tasks, learning paradigms, and tools. It is quite challenging but important to categorize the analytics and learning 'family'. This section attempts to draw a conceptual map of the analytics and learning-centered data science discipline.

There are different types of data inputs, problems, analytical/learning tasks, experiences, evaluation methods, and outputs (e.g., learning a signal or feedback from data or experience) in data science. Analytics and learning approaches can thus

be categorized in terms of their foundations, learning tasks, methods, and business problems.

Figure 7.2 summarizes the main methods, tasks and objectives in machine learning, knowledge discovery, and general data analytics, and the foundations of analytics and learning, and enabling techniques for data science. They can be categorized into the following major groups of techniques:

- Foundations of data science in particular analytics and learning: these include theoretical foundations and tools for analytics and learning in such areas as algebra, numerical computation, set theory, geometry, statistical theory, probability theory, graph theory, and information theory.
- Classic research on analytics and learning: which consists of such areas as feature engineering, dimensionality reduction, rule learning, classic neural networks, statistical learning, evolutionary learning, unsupervised learning, supervised learning, semi-supervised learning, and ensemble learning.
- Advanced research on analytics and learning: which includes such areas as representation learning, Bayesian networks, kernel machines, graphical modeling, reinforcement learning, deep learning (deep neural networks and deep modeling), transfer learning, non-IID learning, X-analytics, advanced techniques for optimization, inference and regularization, and actionable knowledge discovery.
- Enabling techniques for data science: these include artificial intelligence techniques, intelligent systems, intelligent manufacturing techniques, big data and cloud computing techniques, data engineering techniques, Internet of Things techniques, and security techniques.

In the following sections, the main focus is on categorizing and summarizing relevant techniques in the main categories of the data science discipline. More details about classic and advanced techniques for analytics and learning are available in book [67].

7.3.1 Foundations of Data Science

In addition to the broad discussion on multiple disciplinary techniques for data science in Chap. 6, Fig. 7.3 further highlights the fundamentals for analytics and learning. The following aspects of fundamentals are listed, which are commonly required in many analytics and learning tasks and approaches.

- *Algebra*: involves foundations in linear algebra, abstract algebra, group theory, field theory, measure theory, and logic. These include mathematical tools for processing matrices, tensors, norms, eigendecomposition, algebraic structures (such as fields and vector spaces), interaction patterns between objects and their environment, and propositional logic.
- *Numerical computation*: consists of foundations and tools to support the processing of value functions, interpolation, extrapolation and regression, equations,

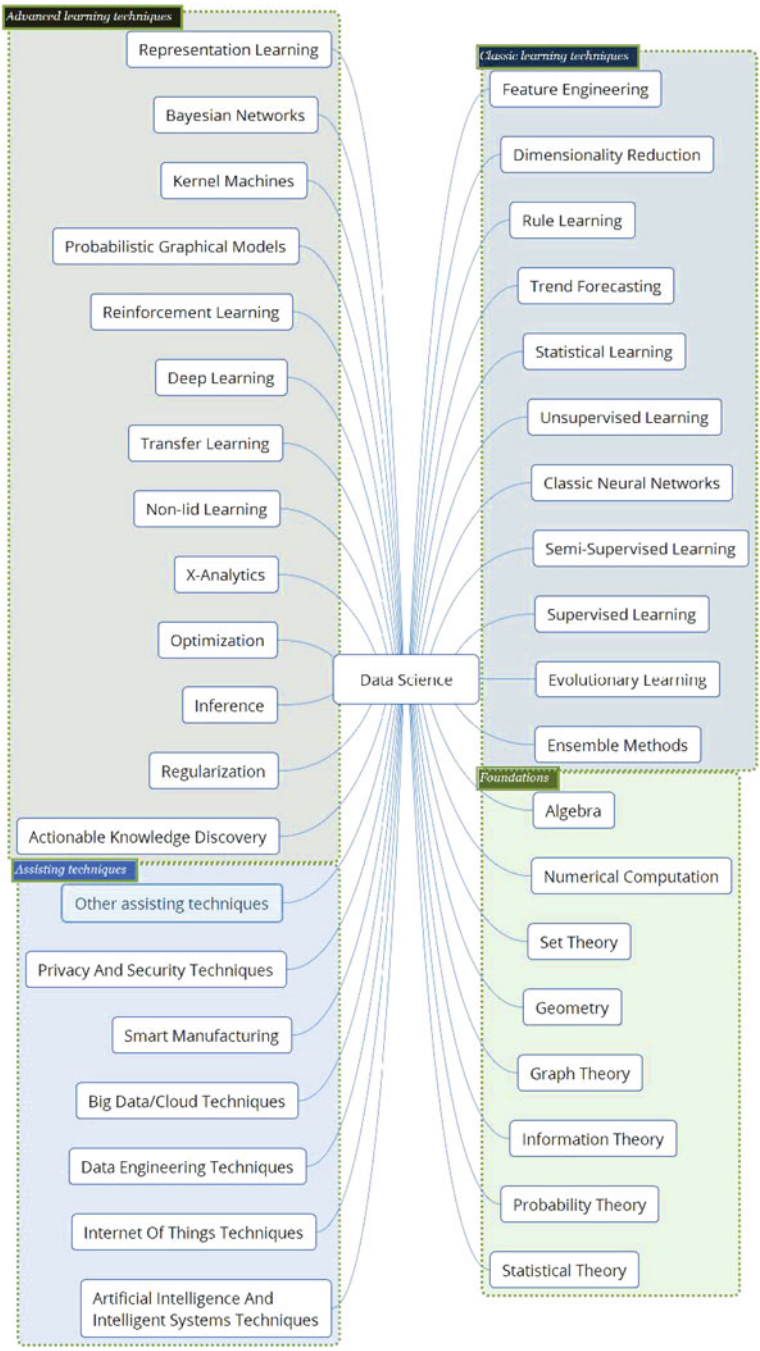

Fig. 7.2 Overview of data science-related techniques

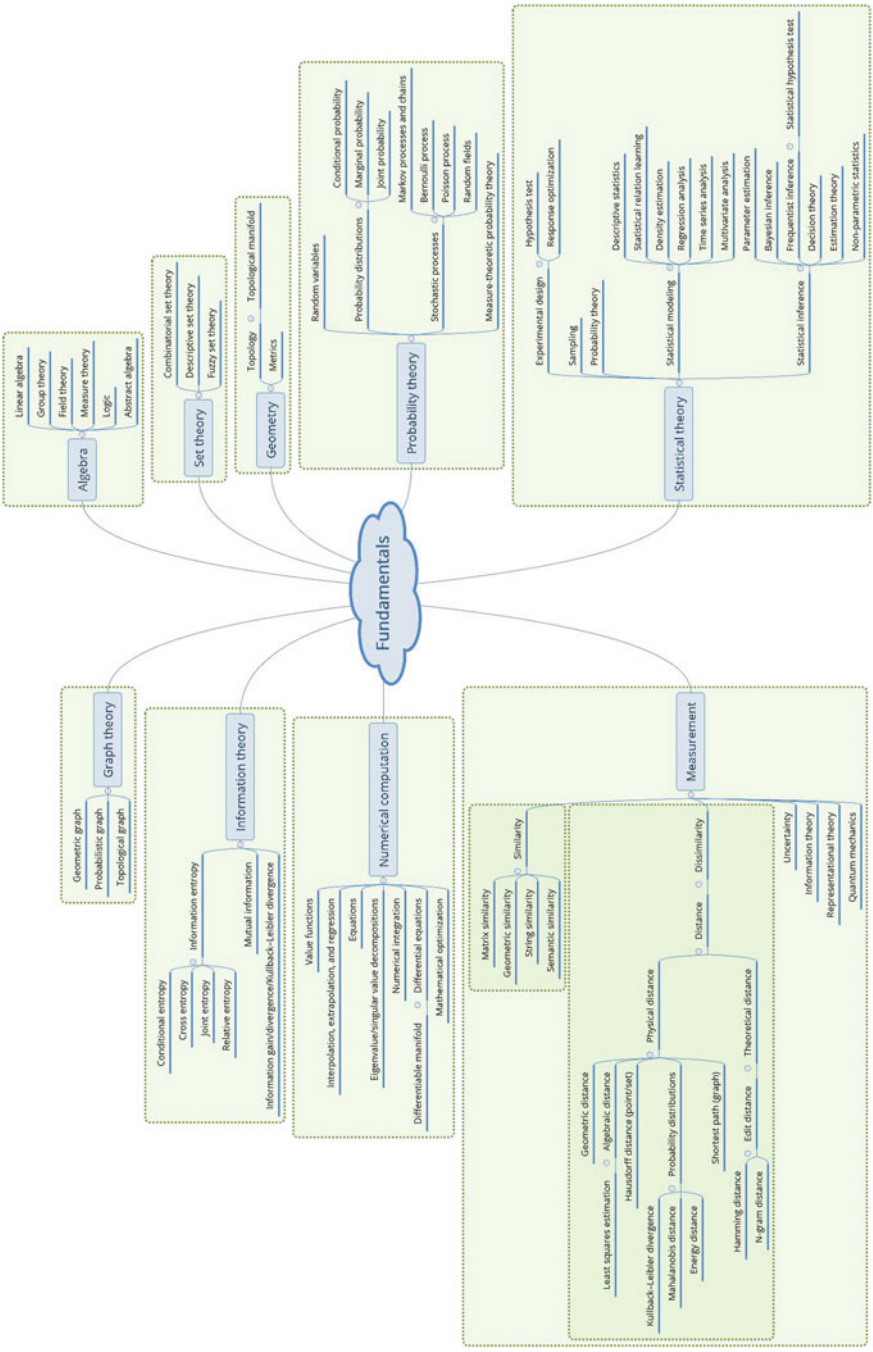

Fig. 7.3 The foundations of analytics and learning

eigenvalue decomposition, singular value decomposition, numerical integration, differential equations, and mathematical optimization.
- *Set theory*: forms the foundation of combinatorial set theory, descriptive set theory, and fuzzy set theory.
- *Geometry*: enables the analytics and learning approaches that are built on topology and geometrics, which may involve algebraic topology such as homotopy groups, homology and cohomology, and geometric topology for manifolds and maps.
- *Graph theory*: consists of geometric graph theory, probabilistic graph theory, and topological graph theory for representing, analyzing, and processing pairwise directed and/or undirected relations, structures, and hierarchies between objects and problems.
- *Information theory*: consists of theoretic foundations for the quantification and communication of information by creating measures such as information entropy (including conditional entropy, cross entropy, joint entropy and relative entropy), mutual information, and information gain (Kullback–Leibler divergence).
- *Probability theory*: contributes to the foundations for handling random variables, probability distributions for conditional, marginal and joint probability, stochastic processes including Markov processes and chains, Bernoulli process, Poisson process and random fields, and measure-theoretic probability theory.
- *Statistical theory*: is the core foundation of analytics and learning. Statistics contribute to the design of experiments for hypothesis testing and response optimization, sampling methods, probability theory, statistical modeling (including descriptive statistics, statistical relation learning, density estimation, regression analysis, time series analysis, and multivariate analysis), and statistical inference (consisting of Bayesian inference, frequentist inference, decision theory, estimation theory, and non-parametric statistics).

More extensive discussion about the inter- and trans-disciplinary foundations of data science and their roles and interactions with data science are available in Chap. 6. In [361], relevant mathematical tools, including algebra, set theory, partial orders, and combinatorics, are introduced. Multivariate data analysis can be found in [232]. In [188], a brief introduction to linear algebra, probability, information theory, numerical computation, and basic machine learning concepts are given. The book by [280] introduces algorithms for information theory, inference and learning. Graph theory can be found in [124]. An extended introduction to statistics and computational statistics is available in [170, 178]. Relevant knowledge about statistics for machine learning can be found in [21].

7.3.2 Classic Analytics and Learning Techniques

In the past three decades or so, many techniques have been developed for analytics and learning, which we called classic techniques. They cover feature engineering,

dimensionality reduction, rule learning, classic neural networks, statistical learning, evolutionary learning, unsupervised learning, supervised learning, semi-supervised learning, and ensemble learning. We briefly summarize these techniques below.

Feature engineering refers to the process and techniques for extracting, constructing, mining, selecting and enhancing features that make data analytics and learning effective and efficient. A *feature* is a term that is often interchangeable with such terms as *attribute* and *variable*, which are the preferred terminologies in some disciplines. Each attribute (variable) captures a characteristic of the underlying data and/or problem. Typical issues in feature engineering include the analysis and processing of feature relevance, discrimination, noise, redundancy, bias, and explosion. These may involve the additional challenges of handling the hierarchy (subspace), structure, dependence, and dimensionality.

Dimensionality reduction is the process of reducing the number of input features by mapping them into a lower-dimensional feature space that is more effective and efficient for handling the curse of dimensionality problem. *High dimensionality* may become a serious issue in almost all learning problems and tasks, particularly in feature engineering, unsupervised learning, supervised learning, and optimization. Typical dimensionality reduction techniques include feature selection and extraction, data mapping and scaling, discriminant analysis, value decomposition, regression analysis, and relation analysis.

Rule learning refers to a general analytics and learning method that identifies, extracts, learns or evolves rules from a set of observations. A *rule* may be represented in terms of a structure like 'antecedent' $\rightarrow$ 'consequence' where $\rightarrow$ implies or co-occurs (or "IF 'condition' THEN 'result"'), to represent, present or apply knowledge from data. In particular, *rule induction* extracts formal rules from data which represent full or partial patterns in the data. Typical tasks and methods for rule learning include association discovery (including association rule mining—or frequent itemset mining—and frequent sequence analysis), learning classifier systems, and artificial immune systems, as well as a range of paradigms such as horn clause induction, rough set rules, inductive logic programming, and version spaces for inducing rules.

Statistical learning refers to a collection of theories and techniques for modeling and understanding complex data, in particular, finding an appropriate predictive function, based on statistical theories and functional analysis [207, 410]. Statistical learning theories form the core foundation of data analytics and machine learning. General modeling and predictive tasks in statistical learning include resampling methods, linear regression, shrinkage approaches, clustering, classification, tree-based methods, and support vector machines.

Unsupervised learning refers to learning processes and approaches for which no labels are given, and for which a learning system estimates the categorization and structures in its input. Unsupervised learning can discover hidden patterns, detect outliers in data, or conduct feature learning. Typical unsupervised learning approaches include clustering, latent variable models, data exploration-based methods, and sparse coding.

Supervised learning refers to the process and techniques for assigning a label to each object by inferring a function from the data with supervision from a "teacher" (i.e., assigned labels on training samples). Typical supervised learning tasks and approaches consist of regression, classification, and ensemble methods.

Forecasting refers to the process and techniques for predicting or estimating the future based on past and present data. Forecasting is important for business management and decision-making, and forms the basis for estimating and planning research and development, capacity, manufacturing, inventory, logistics, manpower, sales and market share, finance and budgeting, and management strategies.

Trend forecasting refers to the prediction or estimation of future trends based on the analysis of past and present data (typically time series data, longitudinal data, and cross-sectional data), which is the main task in forecasting. *Trend* refers to *patternable trends*, such as prevailing tendencies in style or direction, popularity, or seasonal and cyclic behaviors, or *exceptional trends* such as drift or change. Typical trend forecasting tasks and techniques can be categorized into qualitative forecasting, quantitative forecasting, and modern predictive modeling.

Evolutionary learning refers to analytical and learning techniques and processes that are built on the mechanisms inspired by genetic, biological, and natural systems. Typical genetic, biological, and natural systems involve working mechanisms (also called operators) such as selection, cross-over, mutation, recombination, and reproduction. An evolutionary algorithm takes a search heuristic that mimics the process and working mechanisms of natural systems to generate new individuals that better fit and better approximate the given problem-solving solution.

More detailed discussion on these classic techniques is available in the book—Data Science: Techniques and Applications [67], also by the author of this book.

7.3.3 Advanced Analytics and Learning Techniques

There are many recently developed techniques for analytics and learning which we refer to as advanced techniques. Here, we briefly introduce the following: representation learning, kernel methods, Bayesian methods, probabilistic modeling methods, deep learning, reinforcement learning, non-IID learning, transfer learning, actionable knowledge discovery, and optimization techniques.

Representation is a critical process for achieving desirable learning objectives, tasks, results and systems. Representation learning has received significant and increasing attention from many relevant communities including computing, statistics, and learning. The objective of representation learning is to learn latent data, information and knowledge representations. It can be categorized into shallow representation and deep representation, depending on the depth of the representations.

Kernel methods [213, 262] rely on the so-called "kernel trick", in which original features are replaced by a kernel function. Here, *kernel* mathematically denotes a weighting function for a weighted sum or integral of another function. A kernel method involves a user-specified kernel function to convert the raw data

representation to a kernel representation, i.e., in the form of a "feature map" (also called a mapping function) from the raw feature space to the kernel space. A user-specified kernel function can be interpreted as a similarity function that measures the similarity over pairs of data points in the raw data.

Probabilistic graphical models combine probability theory and graph theory in a flexible framework to model a large collection of random variables that are embedded with complex interactions. Probabilistic graphical models such as Bayesian networks and random fields are popularly used, for good reasons, to solve structured prediction problems in a wide variety of application domains including machine learning, bioinformatics, natural language processing, speech recognition, and computer vision. Typical probabilistic graphical models include Bayesian networks and random fields.

Bayesian networks, also called belief networks or directed acyclic graphical models, are probabilistic graphical models. A Bayesian belief network [260, 304] is a probabilistic dependency model that uses Bayesian probabilities to model the dependencies within the knowledge domain; it infers hypotheses by transmitting probabilistic information from data to various hypotheses. A Bayesian belief network is represented as a directed acyclic graph, in which a *node* represents a stochastic variable and an *arc* connecting two nodes represents the Bayesian probabilistic (causal) relationships between the two variables. Each node (variable) may take one of a number of possible states (or values). Here, *belief* measures the certainty of each of these states and refers to the posterior probability of each possible state of a variable, namely the state probabilities, by considering all the available evidence (observations). When the belief in each state of any directly connected node changes, the belief in each state of the node is updated accordingly.

Over the last decade, deep learning has been increasingly recognized for its superior performance when sufficient data and computational power are available. *Deep learning* refers to "a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones." [188] As a result, deep learning "allows computer systems to improve with experience and data". Deep learning research has evolved and been rebranded many times since its first recognition in the artificial intelligence and machine learning communities. Today's deep learning research essentially consists of deep neural networks and deep learning research [188]. Key principles and insights that enable deep neural networks include representation power, hidden power, depth power, backpropagation power, and training + finetuning mechanisms. Deep learning has made significant progress in many areas and applications, including speech recognition, audio and music processing, image and video recognition, multimodal learning, language modeling, natural language processing, information retrieval, and sequence modeling. Representative deep network architectures can be categorized into (1) supervised learning architectures, such as convolutional neural network (CNN) and recurrent neural network; (2) unsupervised learning architectures, such as autoencoder and generative adversarial networks; and (3) hybrid learning architectures, such as the

pre-training + finetuning model DBN+DNN. More discussion on deep learning is available in [188] and in [67].

Reinforcement learning is a general-purpose framework for intelligent problem-solving and decision-making. It had its renaissance in the last decade, migrating from classic reinforcement learning research to advanced reinforcement learning research. In a broad sense and from the understanding and management of learning and decision-making perspectives, *reinforcement learning* refers to computational approaches that enable automated goal-directed decision-making by an individual learning from interactions with its environment. Reinforcement learning concerns making decisions about what to do, i.e., how to map situations to ideal actions, in order to maximize a reward signal to achieve an individual's global and long-term goals. It is about characterizing, solving and optimizing a learning problem, rather than characterizing learning methods and investigating how learning takes place. A learner tries and then discovers which actions ideally impact or produce the expected reward, which may affect subsequent rewards. Typical reinforcement learning methods include dynamic programming, simple Monte Carlo methods, unification of Monte Carlo, temporal-difference learning, and function approximation methods such as artificial neural networks. Advanced reinforcement learning is absorbing significant ideas and supporting tools from other disciplines such as neuroscience, statistics, physics, classic machine learning, and optimization, and emerging areas such as deep learning, advanced analytics and learning, big data technology, intelligence science, and complex systems.

It is often assumed in statistical analysis that data is independent and identically distributed (referred to as IID, or i.i.d.. This IID assumption has been dominantly adopted in analytics and learning, and existing analytical and learning systems have been built on the IID assumption of data and problems. *IID learning* relies on the assumption that all objects are IID, and applies this assumption to objects, object attributes, attribute values, learning objective function determination, and evaluation criteria. However, it has been widely accepted that real-life data and problems are not IID (and are thus called non-IID), i.e., data is non-independent and/or non-identically distributed. In a non-IID data problem, *non-IIDness* refers to any *coupling* and *heterogeneity*. Non-IID learning is the learning system that addresses the non-IIDness in complex data and problems. Non-IID learning applies to almost all analytical and learning tasks and other AI tasks, including memory emulation tasks, analytical, statistical, and learning-based tasks, interaction and recognition tasks, and simulation and optimization.

Transfer learning [321, 420] has been proposed as a general learning methodology for addressing business problems and research issues that involve multiple sources of data (channels, views, databases, etc.) to enhance the analytics and learning in one source by taking advantage of the value in other related data sources. *Transfer learning* is a machine learning methodology that either utilizes the richer and/or labeled data, or transfers the knowledge learned in some domains (called *source domains*) to assist with appropriate learning in other relevant domains (called *target domains*) with significantly limited or less informed (labeled) data. The condition for applying transfer learning is the invariance that exists across domains

and/or tasks. This invariance can be used to associate the source domain/task with the target domain/task, thus the knowledge learned in the source domain/task can be transferred to the target domain/task. The relevant learning methods related to or involved in transfer learning include: learning to learn [383], lifelong machine learning [92], domain adaptation [109, 120], self-taught learning, knowledge transfer, knowledge consolidation, inductive transfer, active learning, context-aware learning (also called context-sensitive learning), multitask learning, metalearning, increment/cumulative learning, one-shot learning, and zero-shot learning. Not all of these are part of the transfer learning family, but there are similarities and differences between each of them and transfer learning.

Actionable knowledge discovery (AKD) addresses the problems of existing analytics and learning and various data characteristics and complexities (as discussed in Sect. 5.5.2), with the goal of achieving data-to-insight-to-decision transformation by data-driven discovery (more on this in Sect. 7.4). *Actionable knowledge* [78, 111] refers to knowledge that informs or enables decision-making actions. The term *actionable knowledge discovery* (AKD) refers to the methodologies, processes, and tools that discover and deliver actionable knowledge from data [57]. AKD undertakes *domain driven data mining* [54, 77] to address the various gaps in classic data mining methodologies and algorithms, in particular, the limited actionability of discovery results. Here, *actionability* [78] refers to the power to work, which is an optimal outcome and objective of AKD through the best integration of six core dimensions: problem, data, environment, model, decision and optimization.

Analytics and learning results often face problems of underfitting and overfitting, which lead to bias and variance in learning capacity. This may be caused by the fact that no models can exactly capture the X-complexities and X-intelligence of data and business problems, thus error minimization and the generalization of models is a critical issue. Appropriate techniques are therefore required to ensure this generalization, which includes regularization, optimization, and sometimes also inference as optimization. While optimization, regularization and inference focus on their respective goals, in practice, there are often connections and complementarity between optimization, regularization and inference. They all aim for a certain common performance (e.g., accuracy, robustness, convergence and efficiency) of algorithms to achieve the desired solutions. The respective methods fit their corresponding problem structures, underlying mechanisms, and performance characteristics. Optimization is often essential for finding the best solutions in analytics and learning. Many techniques and algorithms are available to address the respective optimization problems and conditions. They involve multiple areas of research, including mathematical optimization (in particular numerical analysis and optimization), statistical optimization (in particular risk minimization, approximate optimization), and computational methods (including evolutionary computing, and tricks and mechanisms applicable in learning). Depending on whether constraints are applied on the input values, optimization can be categorized as unconstrained optimization or constrained optimization. In accordance with the domain of inputs, optimization can be categorized as discrete optimization or continuous optimization. The type of objective function and constrained conditions also determines the

optimization method, which can be categorized as convex optimization (including linear optimization, quadratic optimization and cone programming) and nonconvex optimization. In addition, the model developed for optimization may involve different volumes or proportion of samples or have an assumption on the certainty of the feasible domain, leading to deterministic optimization, batch optimization, mini-batch optimization and stochastic optimization. Lastly, many real-life optimization problems are *NP-hard*[1] and cannot be solved with exact solutions in polynomial time. Approximation methods tend to approximate optimal solutions to such problems in polynomial time with provable guarantees on the approximate solutions.

More details about the above techniques are available in [67].

7.3.4 Assisting Techniques

The proper functioning of data science requires many other assisting techniques which are necessary but do not directly handle analytics and learning. Here we briefly introduce the following techniques: artificial intelligence techniques, intelligent systems, intelligent manufacturing techniques, big data and cloud computing techniques, data engineering techniques, Internet of Things techniques, and security techniques.

7.3.4.1 Artificial Intelligence and Intelligent Systems

Artificial intelligence (AI) [344] refers to the intelligence demonstrated by machines rather than humans or other living things. AI research studies intelligent systems that can perceive environment, take actions, and maximize the achievement of goals. Since the first proposal of AI in the 1950s, AI research has experienced several resurgent waves, and has evolved into more and more sub-fields, becoming a major area in computer science and the general IT field. Traditional AI research has focused on developing techniques for perception, reasoning, knowledge representation planning, natural language processing, computational intelligence, neural networks, and search and mathematical optimization.

Although these AI techniques are still used and continue to be developed, today's AI research is mainly grounded on high-performance computing infrastructure, large amounts of data, interdisciplinary theoretical advancement, and advanced analytical and learning systems and algorithms. This can largely be seen in the extensive research and development in big data analytics, large-scale machine learning, computer vision of large visual analysis tasks, deep learning of large image and textual data, and biomedical analytics of large biological and medical data.

[1]NP-hard problems refer to Non-deterministic Polynomial acceptable problems, which indicates that they are at least as hard as the hardest problems in NP.

The current resurgence of AI and intelligent systems is predominantly data science-enabled, and learning and optimization-driven. AI research and intelligent systems have migrated to *data-driven AI*, which discovers and utilizes data intelligence to infer and optimize complex problem understanding and resolution; *human-machine-cooperated AI*, which hybridizes human intelligence and machine intelligence for joint problem-solving of complex problems; and *metasynthetic AI*, which builds on the metasynthesis of various types of intelligence.

Both classic and advanced AI research and techniques can be categorized as symbolic AI, connectionist AI, situated AI, nature-inspired AI, social AI, or metasynthetic AI [62] at all stages of AI research.

- *symbolic AI* which represents *symbolic intelligence* in terms of symbols and logic reasoning.
- *connectionist AI* which represents *connectionist intelligence* in terms of connectionism and networking, especially artificial neural networks.
- *situated AI* which represents *situated intelligence* in terms of multi-agent systems and the interactions within a system and between agents and environment.
- *nature-inspired AI* which represents *natural intelligence* in terms of mimicking the working mechanisms in natural, biological and evolutionary systems.
- *social AI* which represents *social intelligence* in terms of social interactions and collective intelligence in problem-solving.
- *Metasynthetic AI* which represents metasynthetic intelligence by synthesizing human intelligence with other ubiquitous intelligence, including data intelligence, behavior intelligence, social intelligence, organizational intelligence, network intelligence, and natural intelligence according to the theory of metasynthetic engineering [62, 333].

Today's AI has been widely and deeply employed in many existing and emergent business domains and applications, and has demonstrated its significant value for the strategic transformation of existing industries and businesses, and for the generation of new AI-driven businesses. Typical applications are healthcare data and medical imaging data analysis-based applications, driverless car design, cashless e-payment services, sharable businesses (e.g., sharable bikes and cars), smart cities and smart homes, and new-generation financial technology driven by data science and intelligent technologies. In principle, AI and intelligent systems can be applied to almost any domain or purpose, using either more or less data. This application to the economy relies on the integration of intelligent systems, smart manufacturing, big data analytics, e-payment systems, data-driven demand-supply management, and data-driven customer relationship and marketing management.

7.3.4.2 Smart Manufacturing

Smart manufacturing is both the application of data science in manufacturing businesses and an enabling technology for conducting data science in many applications.

Smart manufacturing (or *intelligent manufacturing*) [457] refers to advanced forms of manufacturing that apply sophisticated information technologies to the whole manufacturing process and product life cycle to optimize and advance traditional manufacturing technologies and processes. Here, "smart" indicates that manufacturing enterprises effectively apply advanced intelligent systems to the whole manufacturing ecosystem, rapidly producing new products, dynamically responding to global market supply and demand change, and effecting real-time optimization of manufacturing production and supply-chain systems. Intelligent technologies driving smart manufacturing include digital manufacturing, advanced robotics, industrial Internet of Things and sensor techniques, enterprise application integration, big data processing techniques, industrial connectivity devices and services, rapid prototyping, collaborative virtual factory, advanced human-machine interaction, virtual reality devices, intelligent supply-chain management, and cyber-physical system communication. Data science and artificial intelligence are the main drivers for achieving the so-called Industry 4.0 plan [449].

7.3.4.3 Big Data and Cloud Computing Techniques

Big data [288] is the form of data that is so large and complex that it cannot be processed and analyzed by traditional data processing and analysis theories, infrastructure, algorithms and tools. *Big data* generally refers to voluminous and complex data, and the relevant technologies and systems to acquire, store, transfer, query, share, manage, process, analyze, present, and apply such data. This understanding of *big data technologies* exceeds the multiple V-based definition of big data, in which big data is understood to refer to data described by volume, variety, velocity, veracity, and value. The value of data has to be disclosed by advanced analytics and machine learning, through various X-analytics, building on domain-specific data and problems. Accordingly, big data technologies consist of new-generation technologies and tools for undertaking the above operations on large volumes of complex data.

As distributed and parallel computing has become essential in managing, processing and analyzing big data, *cloud computing* has emerged as a new generation of infrastructure, platforms, services and applications to store, access, share, compute, and analyze big data over the Internet or local area networks. Core technologies to enable cloud computing consist of computing architectures, service models (involving infrastructure, platform, software, and functionalities as services), deployment models (private, public or hybrid cloud), and security and privacy. Cloud computing technologies transform the traditional computing paradigms from enterprise-owned private data and computing to service-based centralized and dedicated data and computing. Cloud computing thus enables more cost-effective operations, more professional management of data and computing, device and location independent access and usage, and more scalable and efficient computing. Typical cloud computing platforms and packages consist of the Apache open source toolset, Amazon Web Services Cloud, Google Cloud platform, Microsoft Azure, and many other

cloud services operated by specific vendors. In addition to tools for implementing general data engineering and management functionalities such as service, storage, access, sharing, query, and management, most cloud platforms involve more or less analytics and machine learning modules and form cloud analytics capabilities.

A typical big data and cloud computing platform is the open source Apache framework and ecosystem based on Hadoop. Hadoop generally refers to a collection of software packages built on the Hadoop framework, which consists of a storage system—Hadoop Distributed File Systems (HDFS), and a processing system—MapReduce programming model. Large data (files) are split (mapped) into blocks and distributed across nodes in a large computing cluster for processing in parallel. The processed results from distributed nodes are summarized (reduced) to form the answers to the original processing task. The Hadoop-based cloud computing ecosystem consists of many software packages to handle different and specific functionalities. Examples in the Apache Software Foundation [10] consist of

- Apache Spark, a cluster-computing architecture for cluster-based distributed storage, programming, cluster management (including task dispatching and scheduling), application programming interface, and fault tolerance.
- Apache Ambari, a Hadoop management web interface for provisioning, managing and monitoring Hadoop clusters.
- Apache HBase, a non-relational distributed database for column-based key-value-oriented storage, compression, and in-memory operation in the Cloud Bigtable form.
- Apache Hive, a Hadoop-based data warehouse for undertaking SQL-like data summarization, query and analysis.
- Apache Pig, a software for programming and executing MapReduce jobs on very large data on the Apache Hadoop framework.
- Apache Zookeeper, a centralized service platform for distributed service configuration, synchronization, and naming registry for large distributed systems
- Apache Sqoop, for data connection and transfer between relational databases and Hadoop.
- Apache Oozie, a workflow management system for handling workflow, scheduling and Hadoop jobs.
- Apache Storm, a distributed stream processing framework for stream processing.

7.3.4.4 Data Engineering Techniques

Data engineering techniques prepare data for analysis and learning. They design, build, manage, and evaluate data-intensive systems data science workload from the engineering aspect. Today's data engineering technologies are seamlessly integrated with big data technologies and cloud computing.

Specifically, *data engineering* provides technologies and solutions for database, data warehouse, hardware and memory storage system architecture, construction, management, privacy and security; data, metadata and application integration and

interoperability; distributed, parallel, high-performance, and peer-to-peer data management; cloud and service computing; data and information extraction, cleaning, retrieval, provenance, workflow, query, indexing, processing, and optimization. These processes may involve different types of data, e.g., strings, text, keywords, streams, temporal data, spatial data, mobile data, multimedia data, graph data, web data, and social networks and social media data.

7.3.4.5 Internet of Things

The Internet of Things (IoT) [450] refers to the ecosystem that connects physical devices, home appliances, vehicles, wearable equipment, and other equipment to create a network via the Internet, wireless networks, or other networking infrastructure for exchanging data, remotely sensing or controlling objects in the network. IoT techniques connect physical systems with virtual cyberspace to form cyber-physical systems, which can essentially connect with and integrate the Internet (including mobile networks and social networks) with everything in the physical world, e.g., homes, cities, factories, transport systems, mobile vehicles, humans, and other living entities or objects.

Enabling technologies for IoT consist of IoT network infrastructure; data, information, application and device integration and fusion; wireless networks; sensors and sensor networks (including RFID technology); data storage, access sharing and connection; object remote access, connection, and management; remote locating, addressing, and accessing; tele-operations and remote control; and data and object security, surveillance and privacy.

IoT techniques involve a growing number of applications, e.g., smart homes, smart cities, smart grids, intelligent transportation, and industrial domains (referred to as industrial IoT). Data science plays a critical role in IoT for intelligent analysis, alerting, prediction, optimization, and the control and intervention of devices on the IoT. IoT also expands data science to traditional non-digitized things and areas and enables their data-driven transformation.

7.3.4.6 Security and Privacy Protection Techniques

As discussed in Sect. 4.8.1, data privacy and security are important assurance aspects of data science, technology and economy. They may incur severe ethical, social, economic, legal or political challenges if they are not effectively protected.

Privacy and security protection technologies consist of software, hardware, management and governance, and regulation for preserving and enhancing the privacy and security of devices, systems, networks, and to protect networking, sharing, messaging, communications, and payments. Effective and efficient techniques and tools are required to predict, detect, prevent and intervene in risky behaviors and events. Other mechanisms and actions for preserving and enhancing privacy and

security include creating laws and policies for data protection and authentication, as well as active and automated regulation and compliance management.

Sufficient privacy and security protection will ensure the healthy development of data science, technology and economy. This becomes increasingly important at a time when anti-security and anti-privacy are becoming professions, and as professional bodies and individuals utilize and develop increasingly advanced anti-security and anti-privacy technologies and tools.

7.3.4.7 Other Assisting Techniques

Many other technologies and tools may be required to implement data science, especially when data science is conducted within a specific domain. Some such assisting techniques and tools may already exist; others will need to be developed to support domain-specific data science tasks. For example, to undertake data science innovation in traditional manufacturing businesses, existing production devices, production lines, factories, workflows, processes, scheduling, management, governance, and communications may have to be digitized, networked, quantified, and automated. This revolution in manufacturing will require the development of corresponding sensors; virtual and networkable devices, lines and factories; intelligent and networked enterprise resource planning systems, management information systems, messaging and communication systems, and risk and case management systems. Other examples include digitized and networked payment systems, supply chain management, demand-supply management systems, dispatching and scheduling systems, identification, authorization and certification systems, and risk and security management systems.

7.4 Data-to-Insight-to-Decision Analytics and Learning

As shown in Fig. 7.4, *data-to-insight-to-decision transfer* has taken place at different times and in different analytic stages along the *whole-of-life span of analytics*. This can also be represented in terms of a range of analytics goals (G) and approaches (A) designed to achieve the *data-to-decision* goal in different data periods—past, present, and future—and for the purpose of generating actionable decisions along the timeline of conducting analytics and during the life of analytics.

- Data-to-Insight: *Data-to-insight* transfer seeks to form a valuable and deep understanding of the true nature, inner character or underlying truths about data and business, which usually involve hidden data complexities and/or business dynamics and problems and their nature and solutions, but are interesting to be deeply, richly and precisely explored.
- Data-to-Decision: The aim of *data-to-decision* transfer is to discover, define and recommend decisions that are informed, supported and verified by data despite

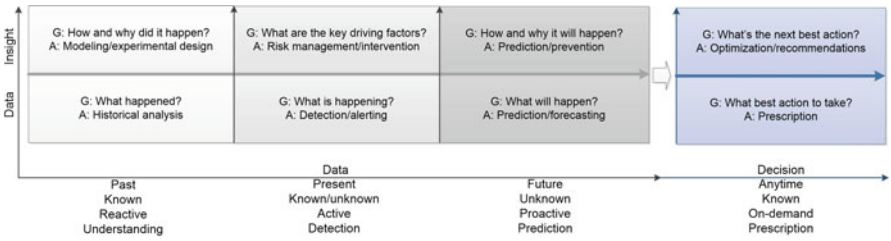

Fig. 7.4 Data-to-insight-to-decision whole-of-life analytics

being invisible. Such decisions are valuable for taking action to achieve better business performance and benefits.

7.4.1 Past Data Analytics and Learning

First, for *past data*, we aim to understand stories that are known to have taken place in the historical data and business; the analytical findings are used for business in a reactive way (here *reactive analytics* means the findings and insights are identified after the event to understand and tackle the problems that have taken place).

The goals and approaches to past data are as follows.

- Goal: The main purpose in past data analytics is to explore "what happened" in the data and business;
- Approach: Historical data analytics methods are used to explore what happened.

Accordingly, the insight extraction from past data is as follows.

- Goal: The purpose is to gain insights into "how and why it happened";
- Approach: Data modeling, experimental design and hypothesis testing are typically applied to utilize the insights gained by this approach.

By understanding past data, this stage is able to focus on "we know what we know" to gain a reactive understanding of what took place and to take appropriate action.

7.4.2 Present Data Analytics and Learning

Second, for *present data*, we are interested in both known and unknown stories in the data and business. Active understanding of hidden data indication is conducted, to detect what is happening in business.

The goals and approaches for present data are:

- Goal: The aim at this stage is to explore "what is happening";
- Approach: Typical methods such as real-time and current detection and alerting are used to understand "what is happening".

The insight discovery through present data analytics is as follows.

- Goal: The purpose is to generate insights about "how and why it happens";
- Approach: Applying the insights is typically dependent on risk management and intervention.

This stage typically addresses "we know what we do not know", with alerts generated about suspicious events, or interesting groups or patterns presented in the data and business. The insights are extracted for decision-making purposes, such as real-time risk management and intervention, to address the question "what are the key driving factors?"

Both past and present data are usually involved in real-time detection and analysis.

7.4.3 Future Data Analytics and Learning

Third, for *future data*, we focus on mainly unknown stories in data and business. This approach relies on proactive and predictive analytics and data modeling.

On present and future data, we are concerned with the following.

- Goal: To investigate "what will happen" in the future;
- Approach: Typical methods for predictive modeling and forecasting are focused to understand "what will happen".

The insight gained from future data analytics has the following objectives.

- Goal: To gain a deep understanding of "how and why it will happen";
- Approach: The extracted insights are typically used for prediction and prevention of future unexpectedness.

This stage gains a deep understanding of the problem that "we do not know what we do not know" and seeks to find solutions to by estimating the occurrence of future events, grouping and patterns, and undertaking and achieving proactive understanding, forecasting and prediction, and early prevention.

When necessary, data from the past, present and future may all be involved.

7.4.4 Actionable Decision Discovery and Delivery

Lastly, our focus is on consolidating anytime data to recommend and take actions for *actionable decision-making*. At this stage, we gain a fairly solid understanding

of data and business problems, and on-demand mixed analytics are undertaken to provide prescriptive actions for business.

At this stage of data understanding and analytics, we are concerned with the following.

- Goal: To investigate "what is the best action to take" in business;
- Approach: Typical methods may include simulation, experimental design, behavior informatics, and prescriptive modeling to identify the next most appropriate actions (often called *next-best* actions).

Correspondingly, we look for insights that aim for the following.

- Goal: To understand and detect "what is the next best action";
- Approach: The next most appropriate actions to take are applied to optimization and recommendation to improve business.

Prescriptive analytics and actionable knowledge delivery thus interpret findings from past, present and future data, and enable the corresponding optimal actions and recommendations to be put in place based on the findings. This addresses the problem of "how to actively and optimally manage the problems identified" by making optimal recommendations and carrying out actionable interventions.

In practice, as shown in Fig. 7.4, real-life analytics involves past data analytics, present data analytics, and future data analytics. Enterprise analytics requires whole-of-life analytics to implement and achieve the data-to-insight-decision transformation.

7.5 Descriptive-to-Predictive-to-Prescriptive Analytics

The paradigm shift from data analysis to data science constitutes the so-called "new paradigm" [209, 305], i.e., data-driven discovery. The history of analytics from an evolutionary perspective spans two main eras—the era of explicit analytics, and the era of implicit analytics [67]. Analytics practices have seen a significant paradigm shift in three major stages:

(1) Stage 1: descriptive analytics and reporting;
(2) Stage 2: predictive analytics and business analytics; and
(3) Stage 3: prescriptive analytics and decision making.

Two other analytical terms: *diagnostic analytics* and *cognitive analytics*, are sometimes used in business, in addition to the above three analytics paradigms:

- Diagnostic analytics: the diagnosis of causes, or "why it happened". This function is often included in descriptive analytics, however.
- Cognitive analytics: the act of causing something to happen. This function is often embedded in prescriptive analytics.

Figure 7.5 shows the analytics paradigm shift.

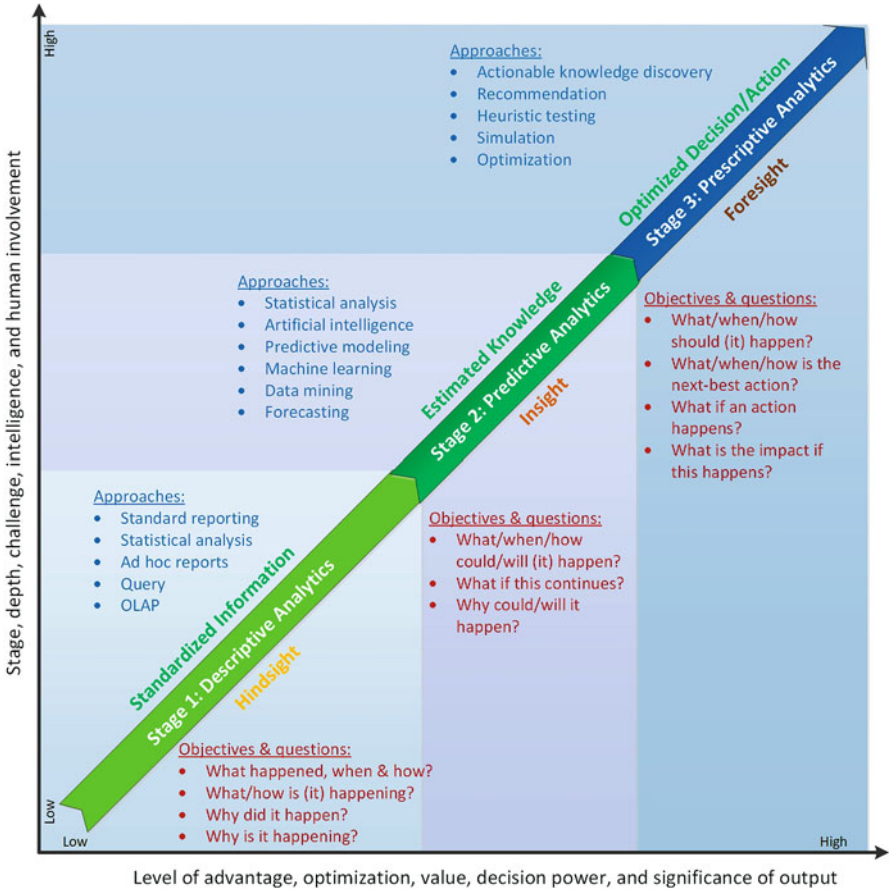

Fig. 7.5 Three stages of analytics: descriptive-to-predictive-to-prescriptive analytics

We briefly discuss the objectives, main approaches and benefits of the above three stages below.

7.5.1 Stage 1: Descriptive Analytics and Business Reporting

Descriptive analytics is the preliminary stage of advanced analytics. *Descriptive analytics* summarizes what has happened or is happening, and characterizes the main features in business.

The main objectives and goals at this stage are:

- answering the questions "what happened?" and "why did it happen?"
- summarizing the main features and trends in business;

- understanding what has happened or is happening in data and business;
- generating descriptions of business operations, performance, dynamics, trends, and exceptional scenarios, with implications about possible driving factors and reasons; and
- presenting regular and periodic summaries and statistics of business concerns.

The main reporting and analytical approaches and methods at this stage include:

- major effort on explicit analytics and standard reporting;
- statistical analysis of data indicators related to or reflecting business operations, performance, dynamics, trends and exceptions;
- generating standard, ad hoc and/or cubic (OLAP-based) reports as periodical summaries and statistics in relation to business;
- querying, drilling up/down;
- factor analysis, correlation analysis, trends analysis and regression; and
- identifying and generating alerts for business management and decision-making.

The consequences and benefits of descriptive analytics for business are:

- to summarize and describe hindsight about business;
- to understand routine and periodical business characteristics, trends and evolution;
- to identify factors and correlations driving the routine and periodical normal development of business;
- to detect exceptional and risky events, occasions, areas and scenarios and identify their implications beyond routine and regular development and trends.

Business reports (often standard analytical reports) generated by dashboards and automated processes are the means for carrying findings from analytics to management.

7.5.2 Stage 2: Predictive Analytics/Learning and Business Analytics

Predictive analytics are a key element of advanced analytics which make predictions about unknown future events, occasions, or outcomes.

The main objectives and goals at this stage are:

- answering the questions "what could/is likely to happen?" and "why could/will it happen?"
- generating a reasonable estimation of the trends and future directions of business;
- predicting what will happen and why it will happen in business;

- estimating the future development of business operations, performance, dynamics, and exceptional scenarios, with implications about possible driving factors and reasons; and
- identifying probable future risk and exceptions.

The main analytical and learning approaches and methods at this stage include:

- forecasting, in particular focusing on regression analysis techniques (including linear/logistic regression, time series analysis, regression trees, and multivariate regression);
- statistical prediction and forecasting of future business operations, performance, dynamics, trends and exceptions; and
- machine learning, artificial intelligence and data/text mining techniques for predictive modeling (classification methods such as neural networks, support vector machines, Naive Bayes and nearest neighbour-based methods; pattern-based methods; computational intelligence methods such as fuzzy set and evolutionary computing; trend prediction methods; and geospatial predictive modeling).

Predictive analytics may bring benefits and impact to business such as:

- identifying insights about the future;
- understanding the likelihood of future business trends and evolution;
- identifying probable factors that will drive the future development of the business;
- predicting exceptional and risky events, occasions, and scenarios, and determining possible driving factors.

Predictors, patterns, scoring and other findings are created for and presented through dashboards and analytical reports to business managers and decision-makers to understand projected future trends and directions, and the reasons behind them. Analytical reports are delivered to predict trends and exceptions, and to explain the underlying factors.

7.5.3 Stage 3: Prescriptive Analytics and Decision Making

Prescriptive analytics is the most advanced stage of advanced analytics. It suggests the optimal actions to take in decision-making.

The main objectives and goals at this stage are:

- to answer the questions "what should happen?" and "why should it happen?";
- to estimate possible impact, i.e. "what would be the impact if an action is taken?"; and
- to recommend the possible next-best action if a particular action has already been taken, or something untoward happens.

To achieve its objectives, this stage may take the following approaches and methods:

- optimization to generate optimal recommendations;
- simulations to obtain the best possible scenarios and understand the impact of taking action;
- heuristic testing of different options and possible effects; and
- actionable knowledge discovery [57, 59, 77] to ensure that the recommended options are actionable.

Prescriptive analytics contributes directly to business decision-making, and is thus more beneficial for problem-solving and business impact. It achieves this by, for example:

- enabling the immediate application of analytical results by recommending optimal actions to be taken on business change;
- estimating the effect of one option over another, and recommending the better option for prioritized actions;
- promoting outcome-driven analytics and transforming actionable analytics and learning for effective problem-solving and better business outcomes.

Prescriptive decision-taking strategies, business rules, proposed actions and recommendations are subsequently disseminated to decision-makers for the purpose of taking corresponding action.

7.5.4 Focus Shifting Between Analytics/Learning Stages

Figure 7.6 illustrates the shift in focus between three stages of analytics: descriptive analytics to predictive analytics and prescriptive analytics.

At the descriptive analytics and business reporting stage, limited implicit analytics effort is made for hidden knowledge discovery, and even less effort is made in actionable knowledge discovery. Descriptive analytics is mainly (80% of effort) achieved by using off-the-shelf tools and built-in algorithms for explicit analytics.

At the predictive analytics stage, significantly more effort (less than but close to 50%) is made in implicit analytics, which focuses on predictive modeling. Descriptive analytics still plays an important role here, particularly for business analytics. *Business analytics* elevates the process to another level, which aims to gain an in-depth understanding of business through deep analytics. By contrast, classic business analytics, which is widely adopted in business and management, mainly focuses on descriptive analytics. Actionable knowledge discovery receives greater attention, with more effort being made to apply forecasting, data mining and machine learning tools for deep business understanding and prediction.

At the stage of prescriptive analytics, the major (80% plus) effort is focused on the suggestions and delivery of recommended optimal (next best) actions for domain-specific business decisions. This is accompanied by discovering invisible and actionable knowledge and insights from complex data, behavior and environment and by implicit analytics in a specific domain by considering its specific X-complexities. About 50% of the effort is made on implicit analytics and actionable knowledge discovery. As a result, innovative and effective customized algorithms and tools are invented to deeply and genuinely understand domain-specific data and business, and significant effort is expended to discover and deliver actionable knowledge and insights. This forms the scenario of personalized analytics and learning, tailored for specific data, domain, behavior and decision-making. In contrast, relatively limited (less than 20%) effort is expended on explicit analytics, since this is conducted through automated processes and systems.

During the paradigm shift (as shown in Fig. 7.6), a significant decrease is seen in the effort expended on routine explicit analytics, which is increasingly undertaken by automated analytics services. By contrast, a significant increase in effort is seen in implicit analytics and actionable knowledge delivery [77]. The shift from a lower stage to a higher stage accommodates an increasingly higher degree of knowledge, intelligence and value to an organization, but it also means there are more challenges to face.

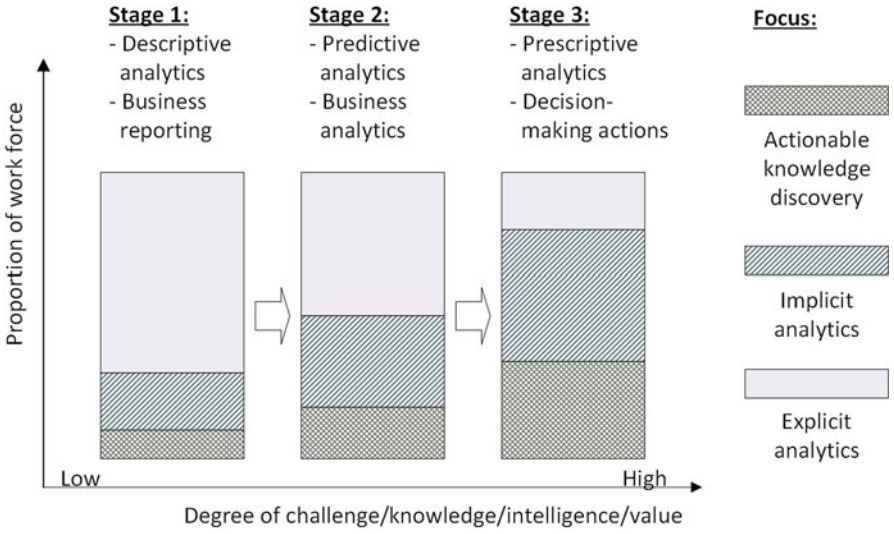

Fig. 7.6 Focus shift from descriptive to predictive and prescriptive analytics

7.5.5 *Synergizing Descriptive, Predictive and Prescriptive Analytics*

Table 7.1 summarizes the key aspects and characteristics of descriptive, predictive and prescriptive analytics.

The summarization and comparison of descriptive analytics, predictive analytics and prescriptive analytics shown in Fig. 7.5 and Table 7.1 clearly show the complementarities between them. The complementary capabilities are embodied in terms of the objectives, challenges and questions to be addressed, data timeliness, approaches and foci, and levels of intelligence and automation in respective stages.

Table 7.1 Descriptive, predictive and prescriptive analytics

Categories	Descriptive analytics	Predictive analytics	Prescriptive analytics
Definition	Summarize what happened or is happening, and characterize the main features	Make predictions about unknown future (events, occasions, or outcomes)	Optimize indications and recommend best actions for smart decision-making
Objectives	Understand what happened or is happening in data and business	Predict what will happen and why it will happen in business	Suggest the next-best actions and estimate possible impact
Challenge	Low to standard	Medium to advanced	Advanced
Questions	What happened, when and how? or What is happening, when and how? Why did it happen or is it happening?	What could/will happen, when and how? What if this continues? Why will it happen next?	What is the next-best action to take, when and how? What happens if an action is taken? What is the impact if it happens?
Data	Past data and/or present data	Past data and/or present data	Past data and/or present data
Timeliness	Historical and/or present/real-time	Future	Future
Approaches	Statistical analysis, standard reports ad hoc reports, dashboards/ scorecards query, OLAP, visualization	Predictive modeling, data/ text mining, forecasting, statistical analysis, artificial intelligence, machine learning	Optimization, simulation, testing, experimental design, heuristics, actionable knowledge discovery
Intelligence	Hindsight	Insight	Foresight
Automation	High	High to standard	Low with human involvement
Advantage	Low to standard	Medium to advanced	Advanced
Output	Reporting, alerting, trends, exceptions, factors and implications	Likelihood of future, unknown events or unknown outcomes	Optimal actions, decisions or interventions, better effect and impact
Decision power	Low	Medium	High

In the real world, enterprise analytics often requires all three-stages of analytics. Many analytics case studies follow the descriptive-to-predictive-to-prescriptive analytics transition, as shown in Fig. 7.5, conducted to address different objectives by different approaches. How can the three-stage analytics be synergized in enterprise analytics practices? The following observations offer some direction in undertaking complex data analytics projects. For simple and specific projects, and scenarios that are well understood, such as the predictive modeling of credit card fraud, direct and targeted analytics may be executed.

At the very beginning of a data analytics project, descriptive analytics may be explored on selected samples or sub-topics of data to understand data characteristics and complexities, as well as the main challenges and the level of those challenges. Predictive analytics are then arranged. In general, prescriptive analytics are arranged as the most advanced and late stage of data science project.

For each stage, a comprehensive understanding of analytics plan, applicable approaches, data manipulation, feature engineering, and evaluation methods need to be developed. It is necessary to understand the differences between stages in terms of analytical functions, platform support, data processing, feature preparation, programming, evaluation, and user interactions.

On one hand, the synergy of conducting three stages of analytics when developing an enterprise analytical application is required to incorporate system-level and unified requirement analysis, analytics infrastructure and architecture selection, analytics programming and project management tool selection and configuration, enterprise and external resource acquisition, sharing, matching, connection and integration, feature engineering (including selection, fusion and construction), data governance, and deployment and case management arrangements.

On the other, as each analytics stage involves diversified tasks and purposes, and addresses different challenges and objectives, the support for individual modeling, infrastructure and programming platform configuration, data preparation, feature engineering, and project management, evaluation, deployment and decision-support arrangements can be quite divided and must be specified and customized. This is in addition to the overall connections and collaborations within an enterprise solution.

Skill and capability sets are complementary to all of the above, thus data analysts with appropriate backgrounds, qualifications, knowledge, skill and experience are required to take on the respective roles and responsibilities. This requires the formation of a collaborative data science team, with suitable team management and project management methodologies, tools and performance evaluation systems.

From the perspective of the management of analytical complexities and the feasibility of application and deployment of respective analytics, a systematic plan is required to determine the business scope and area, analytical depth and width, timeliness and data coverage, as well as the analytical milestones and risk management for each type of analytics. This involves the recruitment of data science teams to fulfil the respective plans, cross-team collaboration, communication, and consolidation.

Lastly, as each type of analytics specializes in addressing different business problems and objectives, the collaborations and communications between respective

analytical stages (or analytics team) and business and management teams will differ; for example, involving different areas and levels of operations and management. Synergizing the three stages of analytics approaches thus requires careful planning and the implementation of plans for business communications and decision-making.

7.6 X-Analytics

The data-oriented and data-driven factors, complexities, intelligences and opportunities in specific domains compose the nature and characteristics of data analytics, and drive the application, evolution and dynamics of data analytics.

7.6.1 X-Analytics Spectrum

The application association of data analytics in specific domains has created the phenomenon of domain-specific analytics. *X-analytics* is the general term for analytics in domain-specific data and domain problems, where *X* refers to a specific domain or area.

As shown in Fig. 7.7, typical *domain-specific X-analytics* consists of

- data type and media-based analytics, such as audio and speech data analysis, transactional data analysis, video data analytics, visual analytics, multimedia analytics, text analysis and document analysis, and web analytics;
- intensively explored core business-oriented analytical areas, such as business analytics (general business and management), risk analytics, financial analytics (including accounting, auditing, banking, insurance, and capital markets), government analytics (including social security, taxation, defense, payment, financial service, border protection, statistics and intellectual properties-based), military analytics, marketing analytics, sales analytics, portfolio analytics, operations analytics, customer analytics, health analytics, medical analytics, biomedical analytics, and software analytics;
- popularly explored new business-oriented analytical areas in recent years, such as social analytics (including social network analysis and social media analysis), e-commerce analytics, behavior analytics, mobile analytics, security analytics and intelligence analytics, location-based analytics, and service analytics;
- in recent years, previously rarely explored business-based analytical areas such as manufacturing analytics, logistics analytics, brain analytics, city utility analytics (including energy, water, and transport), agricultural analytics, security analytics and intelligence analytics, learning analytics and education analytics;
- emergent analytical areas such as behavior analytics, people analytics, work analytics, leave analytics, performance analytics, living analytics, urban analytics, and cultural analytics.

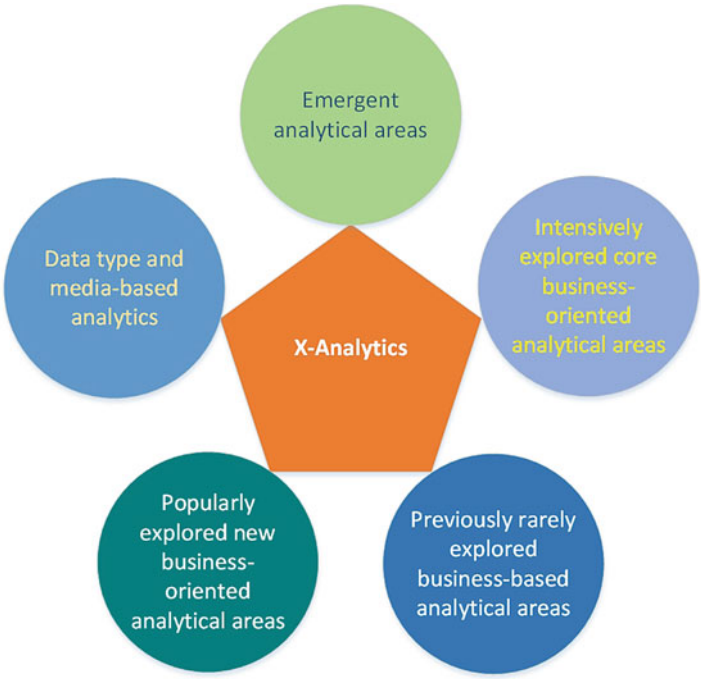

Fig. 7.7 The X-analytics family

These analytical areas and applications largely drive the development, evolution and popularity of data analytics, and the transformation from data analysis to advanced analytics, and from statistics to data science.

7.6.2 X-Analytics Working Mechanism

Typically, a domain-specific X-analytics is formed to address the following aspects:

- understanding and identifying domain-specific complexities and intelligence; e.g., identifying and modeling properties related to behavior in behavior analytics.
- specifying and formalizing domain-specific problems and challenges, and converting them to research issues; e.g., identifying behavior-related problems such as abnormal group behaviors or the impact of negative behaviors, and describing the corresponding behavior-based research issues, such as detecting abnormal group behaviors and quantifying the impact of negative behaviors.

- specifying and constructing domain-specific data structures; e.g., building a behavior model to represent behavior properties and processes, and converting business data to behavioral data.
- identifying research topics and developing analytical theories, models and tools on top of the constructed data structures; e.g., for behavioral data, such topics as behavior pattern mining, group behavior pattern discovery, behavior modeling, and behavior impact modeling may be specified and studied to establish corresponding analytical and learning theories and models.
- developing evaluation systems including measures, processes, and tools to evaluate the performance of relevant research theories, models and results; e.g., behavior utility to quantify the utility of behavior sequences and the significance of specific behaviors.

The process of conducting a domain-specific X-analytics may accordingly include:

- understanding the domain by focusing on specific domain problems and learning relevant domain knowledge;
- identifying and quantifying the specific domain complexities and intelligence;
- converting specific domain complexities to valuable research issues and questions;
- modeling and formalizing domain-oriented problems by proposing corresponding representation structures;
- preparing the data and establishing the corresponding data structures;
- constructing and selecting features and conducting feature analysis and mining;
- representing data features and characteristics;
- developing analytical theories, models and tools to address the identified issues and problems;
- evaluating the performance and impact of the representation and modeling;
- deploying models and developing applications and problem-solving systems.

Figure 7.8 indicates the main research constituents and tasks in establishing and conducting domain-specific X-analytics. In general, X-analytics involves both (1) the top-down modeling and representation of domain factors and issues in terms of holism (see more discussion on holism in Sect. 5.3.1.2) and (2) bottom-up analytical and learning of domain-specific problems and data in terms of reductionism (see more discussion on reductionism in Sect. 5.3.1.1).

7.7 Summary

Analytics and learning are the kernel stone of data science and are ubiquitous in the era of data science and analytics. The ubiquity is embodied in

- various *stages* of analytics and learning, e.g., historical data analysis, real-time detection and forecasting, and future prediction and alerting;

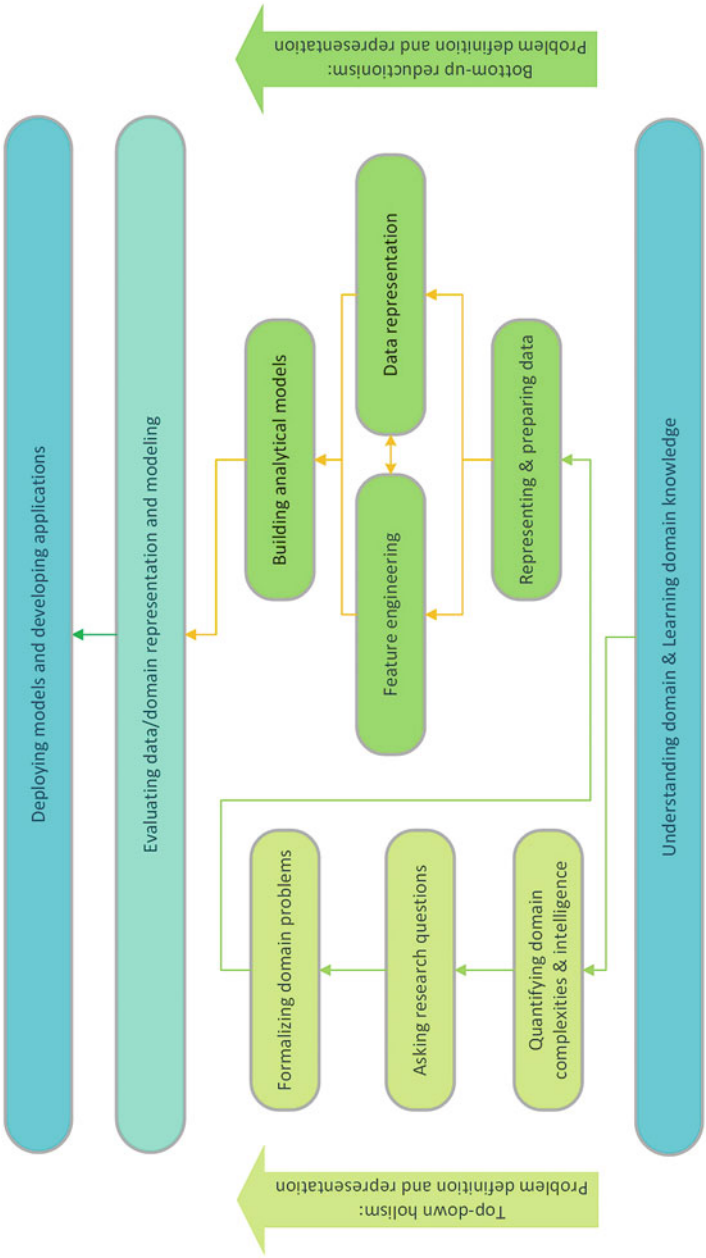

Fig. 7.8 The general X-analytics working mechanism

- diversified *forms* of analytics and learning, e.g., from descriptive analytics to predictive analytics and then to prescriptive analytics;
- a variety of *approaches* to the implementation of analytics and learning, e.g., knowledge-based methods, statistical and probabilistic learning, evolutionary computing-based optimization, and deep networks; and
- different *domains* for undertaking analytics and learning, e.g., financial analytics, government analytics, behavior analytics, and social analytics.

To enable the ubiquitous analytics and learning, different analytics and learning problems, tasks and techniques are required. In book [67], a summary of the family of analytics and learning techniques is provided, following by introduction, categorization and summarization of classic analytics and learning techniques, which cover learning foundations, measurements, and many different classic approaches for conducting diversified analytics and learning problem-solving and tasks. Further, advancements in analytics and learning are presented in the book. The book also introduces the concepts and techniques for inference, optimization and regularization.

The applications of the above discussed analytics and learning are discussed in Chap. 9, which form and drive the data economy and data industrialization, to be discussed in Chap. 8.