Data Science Based on Artificial Intelligence

Chapter · September 2020

DOI: 10.1007/978-3-030-36375-8_1

1 author:

Arthur Kordon
Kordon Consulting LLC
52 PUBLICATIONS **304** CITATIONS

SEE PROFILE

Chapter 1
Data Science Based on Artificial Intelligence

There are three great events in history. One, the creation of the universe. Two, the appearance of life. The third one, which I think is equal in importance, is the appearance of Artificial Intelligence.
Edward Fredkin

Artificial intelligence (AI) represents a paradigm shift that is driving at the same time scientific progress and the evolution of industry. Of special interest to industry is the fast transition of AI from a narrow research area in a limited number of academic labs to a key topic in the business world. This process is combined with an invasion of other new paradigms, such as analytics, big data, and the Internet of Things (IoT). On top of that, a new discipline, named Data Science, has gradually occupied a central place as the leading branch of knowledge that covers all relevant methods and work processes for translating data into actionable business solutions. Learning its basis and key capabilities is becoming a new academic and business challenge. Another critical factor is the appearance of data scientists, who have become the drivers of this transformation.

Unfortunately, this technological avalanche has caught most businesses unprepared. The confusion begins with a bombardment of catchy buzzwords, combined with an invasion of gazillions of vendors offering AI voodoo, and growing anxiety about the future of the business if the new promised opportunities are missed.

The first objective of this chapter is to reduce the confusion by clarifying the key buzzwords and summarizing the big opportunities for the proper use of this package of technologies. The second is to define Data Science based on AI, introducing its key methods and special features. The third is to describe the competitive advantages and challenges of applying AI-based Data Science.

1.1 Big Data, Big Mess, Big Opportunity

The exponential growth of data is a constant challenge in the business world. If properly handled, it opens up new opportunities for competitive advantage and future growth. Of special importance is the current boom in several paradigms related to data, such as big data, analytics, and the Internet of Things (IoT), with the leading role played by AI. Understanding the nature of these technologies and their value creation capabilities is the first step in exploring their great potential.

1.1.1 From Hype to Competitive Advantage

Very often the first introduction to AI and related technologies is driven by hype. On the one hand, this hype contributes to the diffusion of both information and misinformation alike as the technology evolves from the lab to the mainstream. On the other hand, the hype seeds enthusiasm for something new and exciting rising on the technology horizon. Ignorance about the advertised approaches, however, blurs the real assessment required to differentiate hype from reality. As a result, a lot of businesses are hesitant to make decisions about embracing the technology. They need to be convinced with more arguments about the value creation opportunities of AI-driven technologies and their real applicability.

The Hype

Usually when a new technology starts getting noticed, subsequent hype in the industry is inevitable. The hype is partly created by the vendor and consulting communities looking for new business, partly created

by professionals in industry wanting to keep up with the latest trends, and partly created by companies with an ambition to be seen as visionary and to become early champions.

Recently, AI has experienced a "Big Bang" as a prophecy generator. Some examples follow: an AI-Nostradamus claims that "Of all the trends, artificial intelligence is the one that will really impact everything." Another one continues in the same line "You can think of AI like we thought of electricity 100 years ago. When we wanted to improve a process or a procedure, we added electricity. When we wanted to improve the way we made butter, we added electric motors to churns. As a single trend that is going to be overriding, that's pretty much it." A third one concludes "Artificial intelligence is going to have a massive impact on organizations globally."

One medicine for dealing with hype is to consider the famous Amara's law:

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."

An interesting result in accordance with Amara's law is a recent survey based on 588 votes by the popular website KDNuggets.[1] While about 60% of KDnuggets readers think AI and automation will improve society, the optimism drops significantly among those with four or more years of experience in developing AI systems (see Fig. 1.1.) Obviously, the reality correction after gaining experience with AI cools the initial inflated enthusiasm.

.

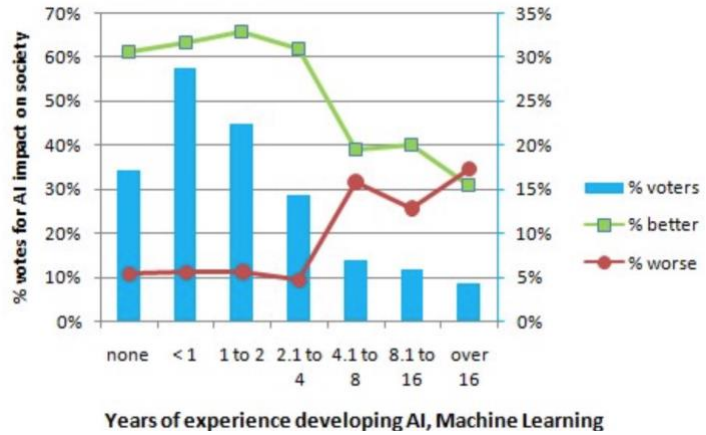


Fig. 1.1 Response to the question "Will AI and automation change society for better or worse?" against years of experience in applying AI.

Another pill against the AI hype is some lessons from the past. At the dawn of AI in 1957, the economist Herbert Simon predicted that computers would beat humans at chess within 10 years. (It took 40.) In 1967 one of the founding fathers of AI, Marvin Minsky, said "Within a generation the problem of creating Artificial Intelligence will be substantially solved." Simon and Minsky were both intellectual giants, but they were bad AI prophets. Thus it is understandable that dramatic claims about future breakthroughs meet with a certain amount of skepticism.

The Mess

When the hype gets out of control, the facts get lost in the haze and the real promise of AI regresses into a high-tech version of snake oil. As a result, companies – including those most likely to benefit from AI-driven solutions – are starting to doubt the claims of AI gurus and vendors promising the Ultimate Intelligence Machine. On the other hand, a side effect of the hype is a growing anxiety about potential lost opportunities. The business-related media amplify these concerns with the image of a ubiquitous AI

[1] https://www.kdnuggets.com/2017/07/optimism-ai-impact-experience.html.

presence in the business of the future. The key recipient of this message is top management. In order to look active and visionary, often the leadership triggers a mess by having large-scale initiatives to introduce AI into their organizations without any preparation. The usual lack of knowledge about AI and related technologies contributes to the mess as well.

One myth that is circulating in the public domain and poisoning the soil for future applications is that AI will cause extensive unemployment and that organizations will need fewer people in the years to come. From current experience with applied AI systems, such systems hardly ever replace an entire job, process, or business model. Most often they complement human activities, which can make people's work even more effective. The best rule for the new division of labor with AI is rarely, if ever, "give all tasks to the machine." Instead, a successful AI implementation of a process will automate some repetitive tasks while it will be more valuable for humans to do the creative tasks.

The first step in reducing this state of confusion is to evaluate the opportunities that AI-based technologies offer, especially those close to the nature of your business.

The Opportunity

Beyond the hype, AI-driven technologies have shown tremendous success in many applications in manufacturing and business. The long list of such applications in manufacturing includes process monitoring and optimization, preventive maintenance, and energy cost reduction. The list of business applications includes price forecasting, customer churn prevention, fraud detection, and human resources optimization. The key application areas will be illustrated with several examples in Chap. 14 for manufacturing and in Chap. 15 for business applications. The most important competitive advantages of AI-driven technologies, such as delivering solutions with "objective intelligence," dealing with uncertainty and complexity, generating novelty, and delivering low-cost modeling and optimization, are discussed later in this chapter. Readers are also recommended to search for the most recent appropriate publicly available AI-driven use cases that can make their business more efficient. Use cases found in this way are the best starting point in the long journey of introducing, exploring, and applying the benefits of AI-driven technologies. The key opportunity hypothesis can be defined as follows: once AI-based systems surpass human performance at a given task, they are much likelier to spread quickly. The first challenge, however, is finding this opportunity in your business.

A key question in estimating the opportunities is the perception of the high cost of these systems. On the hardware side, the concern is that big data and complex AI algorithms will require significant investment. Fortunately, the necessary algorithms and hardware for modern AI can be bought or rented as needed. Google, Amazon, Microsoft, and other companies are making powerful AI-related technology infrastructure available via the cloud. It is assumed that the severe competition among these rivals will cause the prices to drop over time.

Of bigger concern is the cost of gaining knowledge of and learning these new capabilities. The dominant perception is that "AI is rocket science," with very high requirements for math, programming tricks, and bunch of complex algorithms. This is indeed the costliest component of introducing AI-driven technologies. However, AI is not rocket science, and has been implemented in many different places without special requirements for mass-scale training in the technical details. The availability of data scientists, the key drivers of these technologies, is increasing. The training opportunities are exploding, especially online. Training in the era of AI will probably be a continuous process that will gradually include the majority of the labor force. Those who refuse to reskill will surely fail to fit into the new business world order and hence must be ready to miss the boat.

1.1.2 Key Buzzwords Explained

A wide chasm exists between those who build AI-driven technologies and those who use them. Communicating AI using technical jargon will not only confuse potential customers, but also scare them away. The other extreme, of the car-dealer-type pitch, creates a bullshit version of AI that alienates knowledgeable potential users. The best strategy is to explain the new technologies after tailoring one's

conversations to the level of the customers by referencing concepts they already understand, such as brain, network, and tree, and by making AI approachable with friendly examples that clearly demonstrate important use cases.

We would like to follow this strategy in this book, beginning with a condensed explanation of the key buzzwords related to AI-driven technologies.

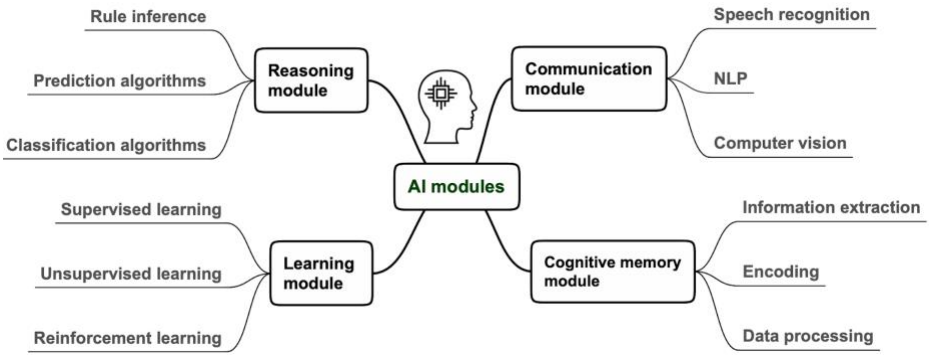
AI

The classical, and still valid, definition of AI, according to *The Handbook of Artificial Intelligence,* is as follows:

> Artificial Intelligence is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior – understanding language, learning reasoning, solving problems, and so on.[2]

The key message is that AI is designed to simulate human thinking. In order to accomplish this goal, an AI system has to interact with the environment, memorize, learn, and reason. A generic structure of an AI system is shown in Fig. 1.2. The following is a short description of the key functional modules:

- *Communication Module.* This includes several methods that mimic human interactions with the outside world through various means – written and spoken language, vision, and signs. One of the key methods is natural language processing (NLP), which identifies the syntax and semantics of written language (a clear example is the Google translator). Spoken language is handled by speech recognition algorithms (Amazon's Alexa is a clear example). Recently, AI capabilities for image processing have grown tremendously, especially in the area of face recognition and self-driving cars.
- *Cognitive memory module.* Similar to the human brain, this module supports the functionality to keep the necessary information for learning and reasoning by the AI system. It includes not only the data but also the parameters of the developed models and the rules of the defined cognitive models.
- *Learning module.* This includes a broad range of algorithms that allow an AI system to learn from available data, knowledge, and a changing environment. The learning can be guided by a teacher (supervised), who prepares training examples in advance and validates the results. The machine can even discover unknown patterns in the data (unsupervised learning) and learn new behavior in a hard way by responding to reward/punishment actions (reinforcement learning). Machine learning capabilities have been available since the late 1980s but recently a novel approach, deep learning, has significantly improved this critical capability of AI systems.



[2] A. Barr and E. Feigenbaum, *The Handbook of Artificial Intelligence*, Morgan Kaufmann, 1981.

Fig. 1.2 Key modules of an AI system

- *Reasoning module***.** This is the least developed module. In the early days of AI, this module included knowledge bases with rules defined by domain experts. This type of AI system is called an expert system and was popular in the late 1980s. However, the reasoning was static and based on subjective knowledge. It is expected that the growing learning capabilities of AI and the new potential of cognitive computing will significantly increase the ways in which machines reason and make decisions.

To summarize, AI is an attempt to make computers as smart, as or even smarter than human beings. It is about giving computers human-like behaviors, thought processes, and reasoning abilities. As a result of this, smart computers enhance human intelligence and increase the potential for value creation.

It is not a surprise that such a complex research area has several different types. One key division into two types, (1) general and (2) narrow AI, is based on the level of generalization. Artificial general intelligence (AGI) – also called strong artificial intelligence – is the intelligence of a machine that can successfully perform any intellectual task that a human being can. Narrow AI – also called weak AI – is designed to perform a narrow task (e.g., only facial recognition, only NLP, or only fault detection). Unfortunately, AGI is still in the front line of research and not ready for industry. The focus in the book is on the narrow AI that has demonstrated enormous capabilities to solve complex problems in a variety of businesses.

Machine Learning

The most widespread AI-driven technology is machine learning. According to Wikipedia, "Machine learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead." 3 The biggest benefit of this technology is that a machine learning algorithm identifies patterns and relationships in data that are used to make predictions about data it has not seen before. It includes several different methods, such as neural networks, support vector machines, random forests, and K-means. Machine learning algorithms are classified based on the desired outcome of the algorithm. The most frequently used algorithm types include supervised learning, unsupervised learning, and reinforcement learning:

- *Supervised learning*. In this, the algorithm generates a function or a classifier that maps inputs to desired outputs. The key assumption is the existence of a "teacher" who provides knowledge about the environment by delivering input–target training samples or labels (Fig. 1.3). The parameters of the learning system are adjusted by reference to the error between the target and the actual response (the output). Supervised learning is the key method in the most popular machine learning approach, neural networks.

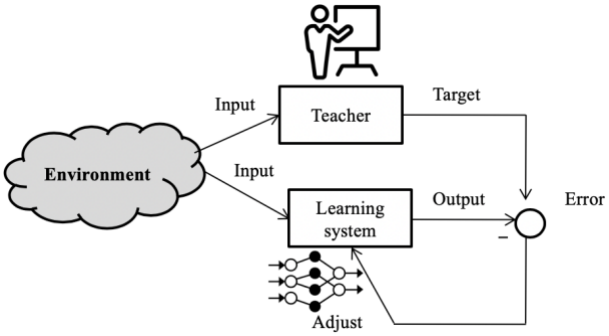


Fig. 1.3 Supervised machine learning

[3] Accessed on June 22, 2019.

Unsupervised learning. The algorithm generates patterns from a set of inputs, since target examples are not available or do not exist at all. Unsupervised learning does not need a teacher and requires the learner to find patterns based on self-organization (Fig. 1.4). The learning system is self-adjusted by the structure discovered in the input data. A typical case of unsupervised learning is that of self-organized maps obtained using neural networks.

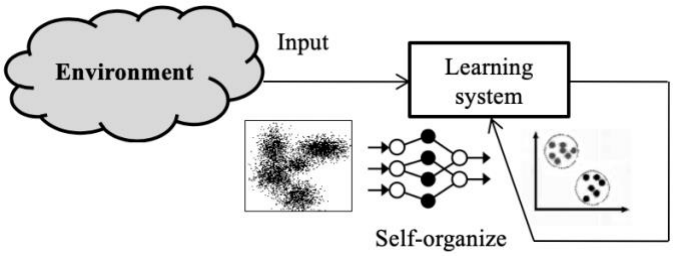


Fig. 1.4 Unsupervised machine learning

- *Reinforcement learning*. The concepts underlying reinforcement learning come from animal behavior studies. One of the most commonly used examples is that of the newborn baby gazelle. Although it is born without any understanding or model of how to use its legs, within minutes it is standing and within 20 minutes it is running. This learning has come from rapidly interacting with its environment, learning which muscle responses are successful, and being rewarded by survival. In the same way the reinforcement learning algorithm is based on the idea of learning by interacting with an environment and adapting one's behavior to maximize an *objective function* specific to this environment (Fig. 1.5). The learning mechanism is based on the trial-and-error of actions and evaluating the reward. Every action has some impact on the environment, and the environment provides a carrot-and-stick type of feedback that guides the learning algorithm. The aim is to find the *optimal behavior*: the one whose actions maximize the long-term reinforcement. Reinforcement learning is often used in intelligent agents and has recently been used in deep learning.

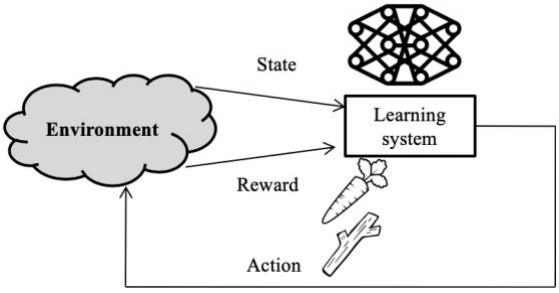


Fig. 1.5 Reinforcement machine learning

Deep Learning

Deep learning is the most sophisticated machine learning approach. The key improvement is that it automates feature selection – the process of learning more abstract representations of the input data (feature extraction) through a sequence of layers This is one of the most time-consuming and critical tasks in developing high-quality models with standard machine learning techniques. Deep learning overcomes this challenge by automatically extracting the most suitable features. In essence, deep learning algorithms learn how to learn.

Usually, a deep learning algorithm is based on a deep neural network. In contrast to the standard neural network with one or two layers, it has many layers of artificial neurons. When a neuron fires, it sends signals to connected neurons in the next layer. During deep learning, connections in the network are strengthened or weakened as needed to improve the performance of the system. An example of a deep learning neural network for image recognition is shown in Fig. 1.6 (Adapted from Fig. 1.2 in http://www.deeplearningbook.org).

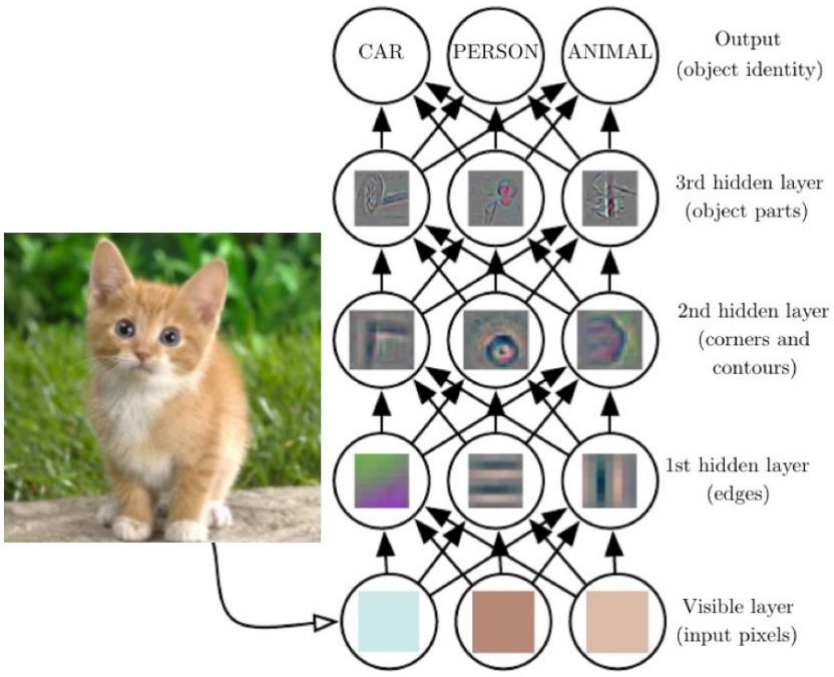


Fig. 1.6 A deep learning neural network structure for image recognition

The original image is broken into pixels, and the first layer detects pixel values. The next hidden layers capture different features in the image. The first hidden layer identifies edges, the second hidden layer recognizes corners and contours, and the third hidden layer identifies object parts that allow the network to classify the object. If a deep neural network is fed pixels of a cat photo, it adjusts its parameters and learns high-level concepts such as "cat." After a deep neural network has learned from thousands of sample cat photos, it can identify cats in new photos as accurately as people can. The "giant leap" from special samples to general concepts during learning gives deep neural networks their power. However, for complex images, the structure of the deep learning neural network becomes very complicated, with more than one hundred layers and millions of tuning parameters. Training such structures takes a lot of time and requires additional hardware, such as graphic processing units (GPUs). Another weakness of deep learning is that generated features are very difficult to interpret.

Big Data

To qualify as big data, data must come into the system at high velocity, with large variation, and at high volumes. Wikipedia says: "Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software."[4] Big is not about the absolute size, but rather about what is necessary to collect, harmonize, store, and analyze the data. The big data paradigm became fashionable due to new data sources such as the IoT, mobile devices, and social media. They created a tremendous growth in the volume, speed, and type of data.

An accepted definition of big data includes the three Vs – Volume, Velocity, and Variety (see Fig. 1.7 for a visualization): [5]

[4] Accessed on June 23, 2019.

[5] Adapted from a post by Michael Walker on 28 November 2012, http://www.datascienceassn.org/blogs/michaelwalker.

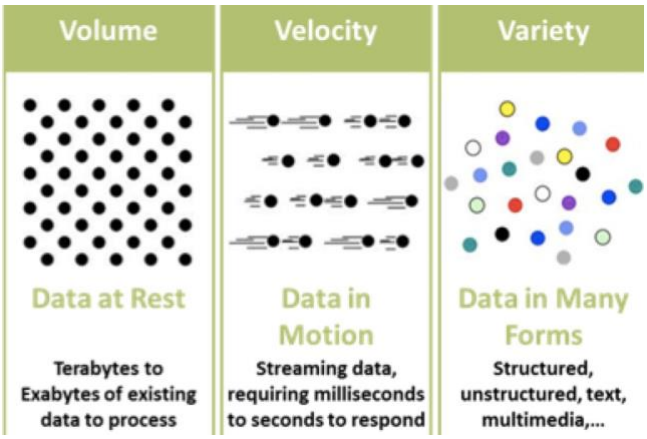


Fig. 1.7 The 3 Vs of big data –Volume, Velocity, and Variety

- *Volume.* The amount of data that organizations need for making decisions has grown tremendously. The first criterion for categorizing data as big is when the size becomes part of the problem. The critical size is business-specific and depends on technological progress. The key business issue is the cost of data collection. The absolute size could be in the range of one petabyte. However, the growing popularity of cloud services has increased significantly the relevant data size and has eased the cost burden.
- *Velocity.* Another feature of recent data is that it is being generated at a much faster rate than data in the past. The growing penetration of the IoT requires data streams in real time.
- *Variety.* Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, pictures, video, and audio.

IoT

The basic idea behind the IoT is that everything consumers and businesses do is now leaving a digital trace – which can be turned into insight and has a potential to create value.

According to Wikipedia, "The Internet of things (IoT) is the extension of Internet connectivity into physical devices and everyday objects. Embedded with electronics, Internet connectivity, and other forms of hardware (such as sensors), these devices can communicate and interact with others over the Internet, and they can be remotely monitored and controlled."[6] Each device or object is uniquely identifiable through its embedded computing system but is able to inter-operate within the existing Internet infrastructure. Experts estimate that the IoT will consist of about 30 billion objects by 2020.

Data being analyzed in near real time provides a lot of valuable IoT use cases. One example is intelligent health-monitoring equipment that can enable faster action than that of doctors or nurses monitoring the same signals manually. Another example is smart homes equipped with smart sensors that can simultaneously increase safety by reducing risks such as fire and flooding, and bring down operational costs by switching heating and air-conditioning on and off at the right times to exploit off-peak rates. There are great expectations that body sensors will track people's activity levels and help change their behavior to improve well-being, while medical sensors can support overall health, for example by monitoring blood sugar levels and dispensing insulin when necessary.

The ongoing advance of AI is causing a natural integration with the IoT. The current trend is toward replacing almost all "dumb" devices with smart machines. Hence the need for AI.

[6] Accessed on June 23, 2019.

Data Analytics

According to Techopedia, "Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements."[7]

This generic buzzword has many specific meanings, the most popular of which have been systematized by Gartner depending on their difficulty and value creation potential as descriptive, diagnostic, predictive, and prescriptive analytics (see Fig. 1.8):[8]

Fig. 1.8 Key types of analytics according to Gartner

- *Descriptive analytics: What happened?* This type of analytics, which is the most used, delivers solutions that help a business to understand better what is going on based on historical data. It provides insight based on patterns and relationships found between the key metrics and measures within the business. Utilizing effective visualization tools enhances the message of descriptive analytics. An example of descriptive analytics is analyzing the key factors and patterns in increasing the energy cost of a business.
- *Diagnostic analytics: Why did it happen?* The purpose of diagnostic analytics is to empower an analyst to drill down and isolate the root cause of a problem. An example of diagnostics analytics is identifying the key factors in and potential sources of equipment failure from available process data.
- *Predictive analytics: What is likely to happen?* Predictive analytics is used to make predictions about unknown future events. Predictive models typically utilize a variety of variables related to a problem to make a forecast. It uses many techniques, such as statistical algorithms, data mining, and machine learning, to analyze available historic data and make predictions about the future. The forecasts are based on discovering trends and patterns from the past that could be valid in the future, as well as identifying the key drivers that influence what we are trying to predict. A typical example of this type of analytics is forecasting of raw materials prices based on selected economic drivers and historical patterns.
- *Prescriptive analytics: What do we need to do?* The highest level of complexity and potential value creation is developing a prescriptive model. This utilizes the insight gained from what has happened, why it has happened, and a variety of "what-might-happen" analysis to help the user determine the course of action to take. Often the recommended actions are derived by sophisticated optimization algorithms that guarantee the best set of decisions for solving a defined problem

[7] Accessed on June 23, 2019.

[8] Adapted from https://www.zdnet.com/article/googling-prescriptive-analytics-youtube-recommendations-and-the-analytics-continuum//.

considering the existing constraints of real-world problems. An example of prescriptive analytics is to use predicted prices in the purchasing of raw materials and minimizing the cost.

Data Mining

Data mining is the process of processing data sets to identify trends, and patterns and discover relationships, to solve business problems or generate new opportunities through the analysis of the data. This discipline is about designing algorithms to extract insights from rather large and potentially unstructured data (for example, text mining). Techniques include pattern recognition, feature selection, clustering, and supervised classification.

Business Intelligence

Business intelligence (BI) is focused on dashboard creation, selection of key performance indicators (KPIs), producing and scheduling data reports based on statistical summaries.

How Are the Buzzwords Related?

An attempt to link the buzzwords discussed above in a systematic way is shown in Fig. 1.9.

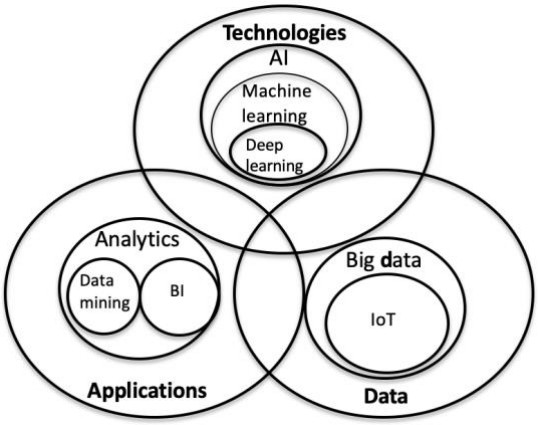


Fig. 1.9 Relationships between key buzzwords

The buzzwords are classified into three categories: Technologies, Applications, and Data. AI is the broadest area in the Technologies. Machine learning is only one of several technologies that are part of AI. Deep learning is part of the machine learning set of algorithms. Analytics is the broadest buzzword in the Applications area. It includes data mining and business intelligence as special steps in the analytics process. Big data is the broadest buzzword in the Data category, and it contains the IoT as one of the key generators of large-volume data in real time.

1.1.3 Why Now?

Several factors are contributing to the recent growing interest in AI-driven approaches. The unique combination of a data and a social media invasion with a tremendous infrastructure growth has opened up new value creation opportunities. A lot of businesses are looking at AI-driven technologies as a key component of their future competitiveness. These factors are discussed briefly below.

Data Invasion

The volume of data generated by digital platforms, wireless sensors, social media, and billions of mobile phones continues to double every three years. It is estimated that the world creates about 2.2 billion gigabytes every day. This data avalanche, however, creates new opportunities for extracting value from unknown patterns and relationships derived from data analysis and modeling. AI-driven approaches are the most powerful tools for translating big data into more insights and more complex relationships. Most of

these technologies, especially machine learning, are data-hungry and need a lot of data to accomplish the learning process. The growth of data is having a very positive effect on the development of AI-driven approaches. It is pushing the research toward novel methods, such as deep learning, that can deliver insight and solutions from large amounts of diverse data such as images, video, and audio.

According to Barry Smyth, a professor of computer science at University College Dublin "Data is to AI what food is to humans."[9] In the same way as different types of food have diverse effects on human body, different qualities of data influence significantly the final result. Just as junk food deteriorates human health, low-quality data leads to models with poor performance.

Social Media Invasion

Social media is more integral to our lives than ever before, and nobody expects it to stop growing anytime soon. Facebook, YouTube, Instagram, and Twitter make it easy for anybody to participate in online communication and online networking, and these platforms have led to a massive global proliferation across demographics. It is expected that the number of users accessing social media will soon surpass 2.5 billion, with 91% of social media users accessing platforms from mobile devices.

Extracting insight and delivering solutions for business use from the available data in social media is becoming one of the key objectives of applied Data Science. AI-driven methods offer the best portfolio of algorithms to fulfill this challenging task.

Infrastructure Growth

Another important factor contributing to the current explosion of AI-driven technologies is the continuous growth of the capabilities of the computing infrastructure. Even as Moore's law[10] is nearing its physical limits, other innovations are fueling this continued progress. Computational power has gotten a boost from an unlikely source: video game consoles. The need for computational power to power even more sophisticated video games has spurred the development of graphics processing units. GPUs have enabled image processing in neural networks that is 10 to 30 times as fast as what conventional CPUs can do.

The growth of cloud-based platforms has given virtually every company the tools and storage capacity to deploy Data Science solutions. Cloud-based storage and analytics services enable even small firms to store their data and process it on distributed servers. Companies can purchase as much disk space as they need, greatly simplifying their data architecture and IT requirements and lowering capital investment. As computation capacity and data storage alike have largely been outsourced, many tools have become accessible, and data can now be more easily combined across sources.

The growth is even more substantial on the software front. The trend of offering free open source software and free versions of popular packages (such as RapidMiner and KNIME) allows AI-based methods to be introduced at a minimal software cost. In parallel, the established vendors, such as SAS, IBM, and Microsoft have updated their key products with improved AI-driven methods that allow development and deployment of complex analytical systems at an enterprise-wide level.

Value Creation Opportunities of AI-Driven Technologies

The value creation opportunities of AI-driven technologies are based on three unique features unlike those of traditional automation solutions. The first feature is their ability to automate complex physical-world tasks that require adaptability and agility. Whereas traditional automation technology is task-specific, the second distinct feature of AI-generated solutions is their ability to solve problems across industries and job titles. The third and most powerful feature of AI-driven technologies is self-learning, enabled by repeatability at scale.

---

[9] http://www.ucd.ie/research/people/computerscience/professorbarrysmyth/.

[10] Moore's law states that the number of transistors included in an integrated circuit doubles approximately every two years.

The self-learning aspect of AI is a fundamental change. Whereas traditional automation capital degrades over time, intelligent automation assets constantly improve.

A significant part of the value obtained from AI-driven technologies will come not from replacing existing labor and capital, but in enabling them to be used much more effectively. For example, AI can enable humans to focus on the parts of their role that add the most value. Also, AI augments labor by complementing human capabilities, offering employees new tools to enhance their natural intelligence.

Another opportunity for value creation from applying AI-driven technologies is their ability to propel innovations. A good example is driverless vehicles. Using a combination of global positioning systems, cameras, computer vision, and machine learning algorithms, driverless cars can enable a machine to sense its surroundings and act accordingly. AI-driven technologies are critical in this process but they allow traditional companies to build new partnerships to stay relevant. As innovation begets innovation, the potential impact of driverless vehicles on economies could eventually extend well beyond the automotive industry. For example, the insurance industry could create new revenue streams from the masses of data that self-driving vehicles generate and the new advanced AI algorithms it uses.

AI-Driven Technologies as a Key Component of Business Competitiveness

Probably the most important shift is that AI-driven technologies, starting from a pure research discipline have become an area of strategic business interest. A recent survey showed that 59% of organizations are still gathering information to build an AI strategy, while 40% are piloting or adopting AI technologies. Tech giants, including Baidu and Google, spent between $20 billion to $30 billion on AI in 2016, with 90% of this spent on R&D and deployment and 10% on AI acquisitions.[11]

The report of that survey cites many examples of internal development, including Amazon's investments in robotics and speech recognition, and Salesforce' investment in intelligent agents and machine learning. BMW, Tesla, and Toyota lead auto manufacturers in their investments in robotics and machine learning for use in driverless cars. Toyota is planning to invest $1 billion in establishing a new research institute devoted to AI for robotics and driverless vehicles.

The competitive advantage of early adopters of AI-driven technologies has already been demonstrated for several sectors in the economy. The results are shown in Fig. 1.10, where the profit margins of companies that apply AI-driven technologies and non-adopters in different sectors of the economy are compared.

[11] McKinsey Global Institute, *Artificial Intelligence: The Next Digital Frontier?* 2017.

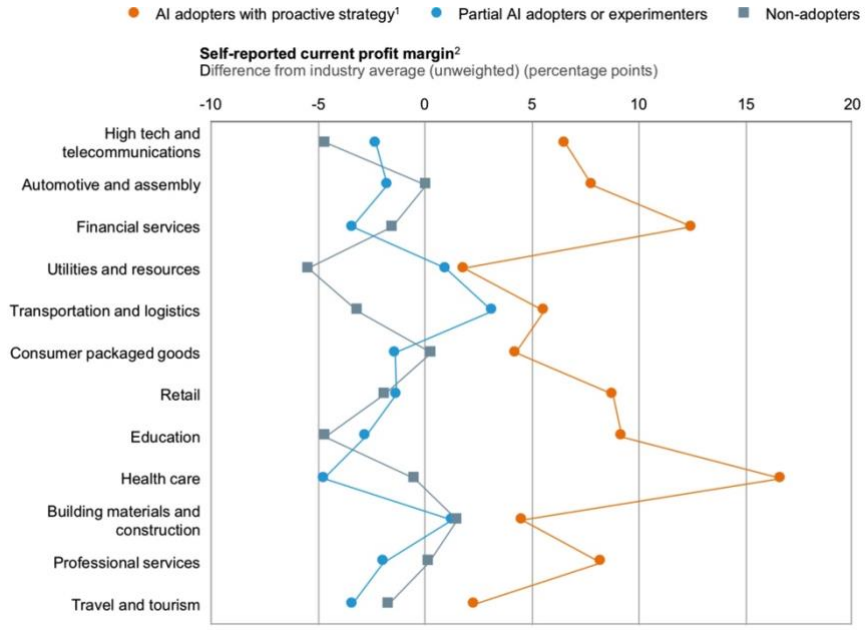


Fig. 1.10 AI early adopters with proactive strategies have higher profit margins

As is shown, healthcare, financial services, and professional services are seeing the greatest increase in their profit margins as a result of AI adoption. McKinsey found that companies that benefit from senior management support for AI initiatives, have invested in infrastructure to support its scale, and have clear business goals achieve a 3 to 15 percentage point higher profit margin than non-AI adopters.

1.2 What Is AI-Based Data Science?

The growing influence of data and AI-driven technologies on science and industry is having an impact on the new rising-star discipline of Data Science. In order to reflect this new reality, a broader version of this discipline, named AI-based Data Science, is proposed in the book. The definition and specific features of AI-based Data Science are discussed below.

1.2.1 Definition of AI-based Data Science

Defining such a broad discipline as Data Science is not an easy task. According to Wikipedia: "Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data."[12]

Data Science is a very complex field that incorporates mathematics, statistics, computer science and programming, math modeling, database technologies, data visualization, data analytics, and so on. From a business point of view, Data Science is a three-legged stool that combines business acumen, data wrangling, and analytics to create high value. Focusing on the hard science skills such as machine learning is not enough. It is one thing to know how to play with fancy algorithms, but it is more important to understand what insights these mathematical models reveal about the business, and what actions to take based on those insights. Experience and business knowledge play a role, as well as curiosity and passion. Sometimes the best results come from nonexperts just because of their desire and persistence.

AI-based Data Science uses AI-driven approaches in addition to statistics to turn data into insight and actions. The specific features of this type of Data Science are discussed below.

[12] Accessed on June 23, 2019.

1.2.2 Features of AI-based Data Science

AI-Focused

The main difference between "standard" Data Science and its AI-based sister is the leading role of AI-driven methods. They add some critical new capabilities to turn data into insight and actions. The key methods of this sort are discussed briefly below:

- *Different knowledge extraction mechanisms*. AI-driven methods offer several options to extract knowledge from available data, such as by using machine learning, simulated evolution, and swarm intelligence. Machine learning algorithms allow a computer to discover patterns and relationships in historical data that could be used for decision-making with unseen data in the future. Simulated evolution automatically generates knowledge from data based on math models fighting with each other. The winners are the best predictors or classifiers derived from the available data. The algorithms of the third option for knowledge extraction from data, swarm intelligence, are based on simulating social interactions among swarms of biological species, for example ants and bees.
- *Decision generation*. Several AI-driven methods, such as decision trees and intelligent agents, allow data to be transformed into automatically generated decisions based on proper quantitative criteria. These decisions can be either executed automatically or used by humans in their decision-making process.
- *Partially automated scientific process*. Several AI-driven approaches, such as machine learning, deep learning, and evolutionary computation, generate their solutions by automatically executing the key steps of the scientific process. In machine learning, the hypothesis is defined either by the labeled data in the case of supervised learning or by automatically discovered patterns (clusters) in the case of unsupervised learning. During the machine learning process, the hypothesis is tested on validation data and a decision is made about its correctness based on model performance.

Broad Set of Technologies

These unique capabilities are based on the new AI-driven technologies. Since AI is an active area of research, the list is continuously growing. Some selected methods will be discussed briefly below, and in more detail in Chap. 3:

- *Machine learning*. The classical machine learning algorithms are based on artificial neural networks, inspired by the capabilities of the brain to process information. A neural network consists of a number of nodes, called neurons, which are a simple mathematical model of the real biological neurons in the brain. The neurons are connected by links, and each link has a numerical weight associated with it. The learned patterns in biological neurons are memorized by the strength of their synaptic links. In a similar way, the learned knowledge in an artificial neural network can be represented by the numerical weights of the mathematical links. In the same way as biological neurons learn new patterns by readjusting the synapse strengths based on positive or negative experience, artificial neural networks learn by readjustment of the numerical weights based on a defined fitness function.
- *Deep learning*. This new technology is also based on neural networks but has very high complexity (sometimes with more than 100 layers and millions of parameters). Deep learning algorithms are used to detect objects in images, analyze sound waves to convert spoken speech to text, or process natural human language into a structured format for analysis. This technology is still in its early days for business use but it is expected to play a key role in the major application area of driverless vehicles.
- *Evolutionary computation*. This automatically generates solutions of a given problem with defined fitness by simulating natural evolution in a computer. Some of the generated solutions have entirely new features, i.e., the technology is capable of creating novelty. In simulated evolution, it is assumed that a fitness function is defined in advance. The process begins with the creation in the computer of a random population of artificial individuals, such as mathematical expressions,

binary strings, symbols, or structures. In each phase of simulated evolution, a new population is created by genetic computer operations, such as mutation, crossover, and copying. As in natural evolution, only the best and the brightest survive and are selected for the next phase. Due to the random nature of simulated evolution, it is repeated several times before the final solutions are selected. Very often the constant fight for high fitness during simulated evolution delivers solutions beyond the existing knowledge about the problem explored.

- *Swarm intelligence*. This explores the advantages of the collective behavior of an artificial flock of computer entities by mimicking the social interactions of animal and human societies. A clear example is the performance of a flock of birds. Of special interest is the behavior of ants, termites, and bees. This approach is a new type of dynamic learning, based on continuous social interchange between the individuals. As a result, swarm intelligence delivers new ways to optimize and classify complex systems in real time. This capability of AI-based Data Science is of special importance for industrial applications in the area of scheduling and control in dynamic environments.
- *Decision trees.* These represent rules, which can be understood by humans and used as knowledge extracted from data. Decision tree output is very easy to interpret even for people from a nonanalytical background. It does not require any statistical knowledge to read and interpret the output. Its graphical representation is very intuitive, and users can easily relate their assumptions.
- *Intelligent agents.* Intelligent agents are artificial entities that have several intelligent features, such as being autonomous, responding appropriately to changes in their environment, persistently pursuing goals, and being flexible, robust, and social by interacting with other agents. Of special importance is the interactive capability of the intelligent agents, since it mimics types of human interaction, such as negotiation, coordination, cooperation, and teamwork. A popular version of intelligent agents, chatbots, are used in machine–human communication.

Usually modern AI is identified by the research and business communities at large with machine learning, and recently has been identified with deep learning. In fact, the whole package of technologies, not only machine learning, is contributing to the current success of AI. One of the objectives of this book is to describe them, demonstrate their applicability, and encourage data scientists to use them.

Broad Applicability

Due to the enhanced capabilities delivered by the broad set of new technologies, AI-based Data Science is highly applicable to many fields, including social media, medicine, security, healthcare, the social sciences, biological sciences, engineering, defense, business, economics, finance, marketing, and many more. Examples of specific applications will be given in Chap. 14 for manufacturing and Chap. 15 for businesses.

Advanced Skillset Required

The new opportunities for improvement that AI-based Data Science offers require additional knowledge about the key AI technologies. The needed skillset includes basic knowledge about the principles of the methods, training in appropriate software platforms of their implementation, and awareness for their application potential.

1.3 Competitive Advantages of AI-Based Data Science

Data Science is becoming a critical factor to maintain competitiveness in the increasingly data-rich business environment. Much like the application of simple statistics, organizations that embrace Data Science will be rewarded, while those that do not will be challenged to keep pace. As more complex, disparate data sets become available, the chasm between these groups will only continue to widen. It is believed that the new, powerful technologies of AI-based Data Science will increase further the competitive advantage of the first group.

The competitive landscape for these technologies includes the following well-established approaches: first-principles modeling, statistics, classical optimization, and heuristics. They are referred to as

competitors in this book. The key competitive advantages of AI-based Data Science are shown in Fig. 1.11 and discussed in detail below.

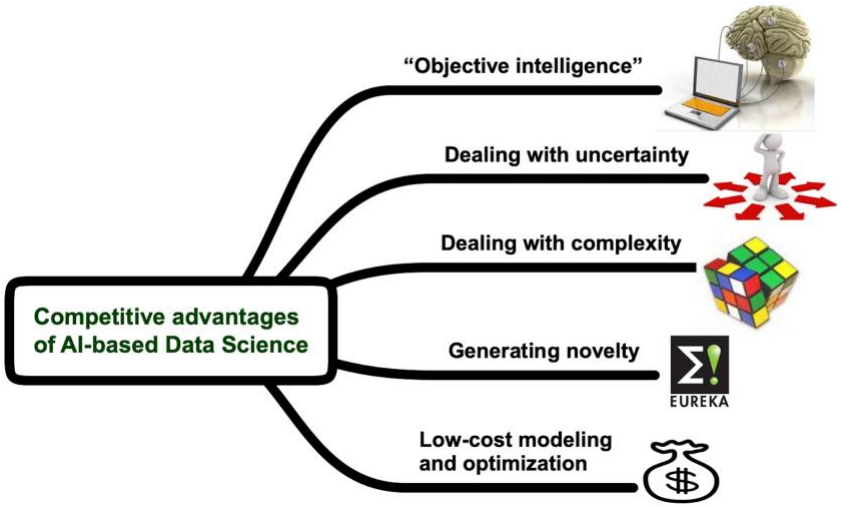


Fig. 1.11. Key competitive advantages of AI-based Data Science

1.3.1 Creating "Objective Intelligence"

The most important feature that boosts AI-based Data Science ahead of the competition is the "objective" nature of the "smart" solutions delivered. "Objective intelligence" is similar to first-principles and statistical models in that those are also "objective," since they are based on the laws of nature and laws of numbers. However, "objective intelligence" is differentiated by its capability to automatically extract solutions from data through machine learning, simulated evolution, or emergent phenomena. In contrast, "subjective intelligence" is based on human assessment only. It is not derived from "objective" sources, such as the laws of nature or empirical dependencies, supported by data. "Subjective intelligence" can be very dangerous if the expertise in a field is scarce. The combination of limited numbers of experts with insufficient knowledge may transform the expected application from a problem solver into a "subjective intelligence" disaster.

The advantages of "objective intelligence" have a significant impact on the application potential of AI-based Data Science. The most important features of "objective intelligence" are discussed below:

- *Consistent decision-making*. The key advantage is that the decisions suggested by "objective intelligence" are derived from and supported by data. As a result, the rules defined are closer to reality and the influence of subjective biases and individual preferences is significantly reduced. Another advantage of "objective intelligence" is that its decisions are not static but adapt to changes in the environment.
- *Nonstop intelligent operation*. "Smart" devices, based on AI-based Data Science, operate continuously and reliably for long periods of time in a wide range of process conditions. As we all know, human intelligence cannot endure a 24/7 mode of intensive intellectual activity. Even the collective intelligence of rotating shifts, typical in manufacturing, has significant fluctuations due to wide differences in the operators' expertise and their attention to the process at a given moment. In contrast, "objective intelligence" continuously refreshes itself by learning from data and knowledge streams. This is one of the key differences between AI-based Data Science and the competition. The competitive solutions can also operate nonstop, but they cannot continuously, without human interference, maintain, update, and enhance their own intelligence.
- *Handling high dimensionality and hidden patterns*. AI-based Data Science can infer solutions from multidimensional spaces with thousands of factors (variables) and millions of records. This feature is beyond the capabilities of human intelligence. Another advantage of "objective intelligence" is its ability

to capture unknown complex patterns from available data. It is extremely difficult for human intelligence to detect patterns with many variables and on different time scales.

- *Continuous self-improvement by learning*. Several learning approaches, such as neural networks, statistical learning theory, and reinforcement learning, are the engines of almost perpetual progress in the capabilities of "objective intelligence." The competitive statistical methods lack this unique feature.
- *No politics*. One can look at "objective intelligence" as an honest and loyal "employee" who works tirelessly to fulfill her/his duties while continuously improving her/his qualifications. Political maneuvering, growing pretensions, and flip-flopping, so typical in the behavior of human intelligence, is unknown. It is not a surprise that this feature sounds very appealing to management.

1.3.2 Dealing with Uncertainty

A key strength of AI-based Data Science is in handling technical uncertainty. The economic benefits of this competitive advantage are substantial. Reduced technical uncertainty leads to tighter control around process quality, faster new product design, and less frequent incidents. All of these benefits explicitly translate technical advantages into value.

One of the advantages of statistics is that uncertainty is built into its foundations. Of special practical importance are statistical estimates of the uncertainty of model predictions, represented by their confidence limits. The different ways in which AI-based Data Science handles uncertainty are discussed below:

- *Minimum a priori assumptions*. Fundamental modeling deals with uncertainty only within strictly defined a priori assumptions, dictated by the validity regions of the laws of nature; statistics handles uncertainty by calculating confidence limits within the ranges of available data; and heuristics explicitly builds the boundaries of validity of the rules. All of these options significantly narrow down the assumption space of the potential solutions and make it very sensitive to changing operating conditions. As a result, their performance lacks robustness and leads to gradually evaporating credibility and imminent death of the application outside the assumption space. In contrast, AI-based Data Science has a very open assumption space and can operate with almost any starting data or pieces of knowledge. The methods that allow AI-based Data Science to operate with minimum a priori information are highlighted below.
- *Reducing uncertainty through learning*. One of the possible ways to deal with unknown operating conditions is through continuous learning. By using several machine learning methods, AI-based Data Science can handle and gradually reduce wide uncertainty. This allows adaptive behavior and low cost.
- *Reducing uncertainty through simulated evolution*. Another approach to fighting unknown conditions is by evolutionary computation. This technology is one of the rare cases when modeling can begin with no a priori assumptions at all. Uncertainty is gradually reduced by the evolving population of potential solutions, and the fittest winners in this process are the final result of this fight with the unknown.
- *Handling uncertainty through self-organization*. In self-organizing systems, such as intelligent agents, new patterns occur spontaneously by interactions, which are internal to the system. As in simulated evolution, this approach operates with no a priori assumptions. Uncertainty is reduced by the new emerging solutions.

1.3.3 Dealing with Complexity

Big data and the Internet of Things have pushed the complexity of real-world applications to levels that were unimaginable even a couple of years ago. A short list includes the following changes: (i) the number of interactive components has risen by several orders of magnitude; (ii) the dynamic environment requires solutions that are capable of both continuous adaptation and abrupt transformations; and (iii) the nature of interactions is depending more and more on time-critical relationships between the components. Another factor that has to be considered in dealing with the increased complexity of practical applications is the required simple dialog with the final user. The growing complexity of the problem and the generated solution must be transparent to the user.

The competitive approaches face significant problems in dealing with complexity. First-principles models have relatively low dimensionality; even statistics has difficulties in dealing with thousands of variables and millions of records; heuristics is very limited in representing large numbers of rules; and classical optimization has computational and convergence problems with complex search spaces of many variables.

The different ways in which AI-based Data Science handles complexity better than the competition are discussed below:

- *Reducing dimensionality through learning*. AI-based Data Science can cluster the data by learning automatically how it is related. This condensed form of information significantly reduces the number of entities representing the system.
- *Reducing complexity through simulated evolution*. Evolutionary computation delivers distilled solutions with low complexity (especially when a complexity measure is included in the fitness function). One side effect of simulated evolution is that the unimportant variables are gradually removed from the final solutions, which leads to automatic variable selection and dimensionality reduction.
- *Handling complex optimization problems*. Evolutionary computation and swarm intelligence may converge and find optimal solutions in noisier and more complex search spaces than the classical approaches can handle.

1.3.4 Generating Novelty

Probably the most valuable competitive advantage of AI-based Data Science is its unique capability to automatically create innovative solutions. In the classical method, before shouting "Eureka," the inventor goes through a broad hypothesis search, and trials of many combinations of different factors. Since the number of hypotheses and factors is close to infinity, the expert also needs "help" from nonscientific forces such as luck, inspiration, a "divine" spark, or even a bathtub or a falling apple. As a result, classical discovery of novelty is an unpredictable process.

AI-based Data Science can increase the chances of success and reduce the overall effort in innovation discovery. Since generating intellectual property is one of the key components of economic competitive advantage, this unique strength of AI-based Data Science may have an enormous economic impact.

The three main ways of generating novelty by AI-based Data Science are discussed next:

- *Capturing emergent phenomena from complex behavior.* Self-organized complex adaptive systems mimic the novelty discovery process by means of their property of *emergence*. This property is a result of coupled interactions between the parts of a system. As a result of these complex interactions, new, unknown patterns emerge. The features of these novel patterns are not inherited or directly derived from any of the parts. Since the emergent phenomena are unpredictable discoveries, they require to be captured, interpreted, and defined by an expert with high imagination.
- *Extracting new structures by simulated evolution*. One specific method in evolutionary computation, genetic programming, can generate almost any types of new structure based on a small number of given building blocks.
- *Finding new relationships*. The most widespread use of AI-based Data Science, however, is in capturing unknown relationships between variables. Of special importance are the complex dependencies derived, which are difficult to reproduce using classical statistics. The development time for finding these relationships is significantly shorter than for building first-principles or statistical models. In the case of simulated evolution, even these dependencies are derived automatically, and the role of the expert is reduced to selection of the most appropriate solutions based on performance and interpretability.

1.3.5 Low-Cost Modeling and Optimization

Finally, what really matters for practical applications is that all the technical competitive advantages of AI-based Data Science discussed above lead to costs of modeling and optimization that are lower than the competition. The key ways in which AI-based Data Science accomplishes those of important advantage are given below:

- *High-quality empirical models*. The models derived by AI-based Data Science, especially by symbolic regression via genetic programming, have optimal accuracy and complexity. On the one hand, they represent adequately the complex dependencies among the influential process variables and deliver accurate predictions. On the other hand, their relatively low complexity allows robust performance in the presence of minor process changes, when most competitive approaches collapse. In general, empirical models have minimal development cost. In addition, high-quality symbolic regression models, with their improved robustness, significantly reduce the deployment and maintenance cost.
- *Optimization for a broad range of operating conditions.* Two popular AI-based Data Science technologies, evolutionary computation and swarm intelligence, broaden the capabilities of classical optimization in conditions with complex surfaces and high dimensionality. As a result, AI-based Data Science gives more technical opportunities to operate with minimal cost in new, previously not optimized areas. Of special importance are the dynamic optimization options of swarm intelligence, where the process could track continuously, in real time, the economic optimum.
- *Low total cost of ownership of modeling and optimization*. All of the advantages discussed above contribute to an overall reduction of the total cost of ownership of the combined modeling and optimization efforts driven by AI-based Data Science. Some competing technologies may have smaller components in the cost. For example, the development and deployment cost of statistics is much lower. However, considering all the components, especially the growing share of maintenance costs in the cost of modeling and optimization, AI-based Data Science is a clear winner. The more complex the problem, the bigger the advantages of using this emerging technology. All known technical competitors have very limited capabilities to handle imprecision, uncertainty, and complexity and to generate novelty. As a result, they operate inadequately in new operating conditions, reducing profit and pushing maintenance costs through the roof.

Still, the biggest issue in estimating the total cost of ownership is the high introductory cost of the technology. Since AI-based Data Science is virtually unknown in industry at large, significant marketing and training efforts are needed. One of the purposes of this book is to suggest solutions that will reduce this cost.

1.4 Key Challenges in Applying AI-Based Data Science

In order to give an objective assessment of AI-based Data Science we need also to identify the potential issues of technical and nontechnical nature.

1.4.1 Technical Issues in Applying AI-Based Data Science

The important technical issues that may reduce the efficiency of AI-based Data Science applications are shown in Fig. 1.12 and discussed next.

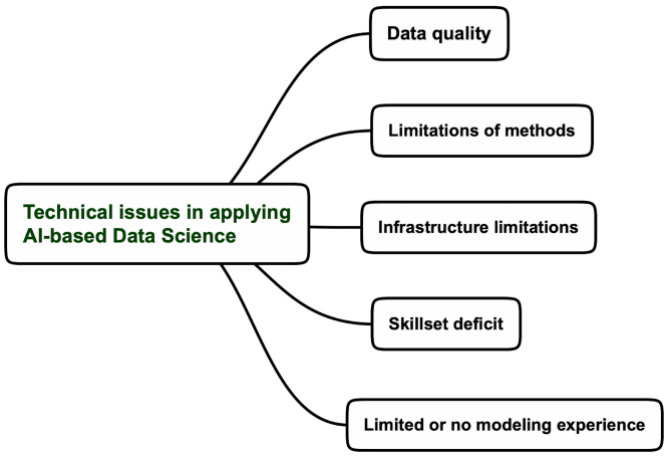


Fig. 1.12 Key technical issues in applying AI-based Data Science

Data Quality

It is not a surprise that data quality becomes a critical factor for the success of AI-based Data Science. Firstly, data availability must be checked very carefully. It is possible that the historical records may be too short to capture seasonal effects or trends. Secondly, the ranges of the most important factors in the data have to be as broad as possible to represent nonlinear behavior. Data-driven models developed on narrow data ranges have low robustness and require frequent readjustment. Thirdly, the frequency of data collection must be adequate for the nature of the modeling. For example, dynamic modeling requires more frequent collection and data sampling. Steady-state models, on the other hand, assume a low data collection frequency that filters out dynamic effects. Fourthly, the noise level has to be within acceptable limits to avoid the classical Garbage-In-Garbage-Out (GIGO) effect. In cases where some of these requirements are not met, it is recommended to create an adequate data collection infrastructure and to begin the application only after collecting the right data. Making a compromise on data quality is one of the most frequent mistakes in applying AI-based Data Science.

Limitations of AI-Based Methods

Each AI-based Data Science method has its own limitations, which will be discussed in detail in Chap. 4. The major issues that these weaknesses create are listed below. Some of these limitations can be compensated by combination with other AI-based Data Science approaches:

- *Black-box models*. Many users view neural networks (the dominant machine learning technology) as magic pieces of software that represent unknown patterns or relationships in the data. The difficult interpretation of the magic, however, creates a problem. The purely mathematical description of even simple neural networks is not easy to understand. A black box links the input parameters to the outputs and does not give any insight into the nature of the relationships. As a result, black boxes are not well accepted by the majority of users, especially in manufacturing, having in mind the big responsibility of controlling plants.
- *Tacit knowledge generation*. In many cases deep neural networks cannot be interpreted at all. Their structure may have tens of millions of connections, each of which contributes a small amount to the developed model. In this case we have a reverse version of the famous Polanyi's paradox: machines know more than they can tell us.[13] Unfortunately, the generated tacit knowledge coded in the layers of deep learning neural net cannot be understood, and it is risky to reuse it. This limits the application of this method in some industries, such as finance and insurance, that require transparency and interpretability of models that are developed and applied.
- *Poor extrapolation*. The excellent approximation capabilities of neural networks within the range of the model development data are not valid when the model operates in unknown process conditions. It is true that empirical models also cannot guarantee reliable predictions outside the initial model development range, defined by the available process data. However, the various empirical modeling methods deliver different levels of degrading performance in unknown process conditions. Unfortunately, neural networks are very sensitive to unknown process changes. The model quality significantly deteriorates even for minor deviations (<10% outside the model development range). A potential solution for improving the extrapolation performance of neural networks is to use evolutionary computation to select the optimal structure of the neural network.
- *Maintenance nightmare*. The combination of poor extrapolation and black box models significantly complicates the maintenance of neural networks. The majority of industrial processes experience changes in their operating conditions of more than 10% during a typical business cycle. As a result, the performance of deployed neural network models degrades, and triggers frequent model retuning, and even complete redesign. Since the maintenance and support of neural networks requires special training, this inevitably increases maintenance cost.

[13] Polanyi's paradox states that "we know more than we can tell, i.e., many of the tasks we perform rely on tacit, intuitive knowledge that is difficult to codify and automate."

- *Computationally intensive.* Deep learning and evolutionary computation require substantial number-crunching power. Fortunately, the continuous growth of computational power according to Moore's law and the new hardware options offered by GPUs and specialized neural net chips are gradually resolving this issue. Improved algorithms are making additional gains in productivity and the companies driving the AI field are continuously developing new, more powerful tools. The third way to reduce computational time is to shrink down the dimensionality of the data by effective variable/feature selection.
- *Time-consuming solution generation.* An inevitable effect of computationally intensive AI-driven methods, such as deep learning and evolutionary computation, is slow model generation. Depending on the dimensionality and the nature of the application, this may take hours, or even days.

Infrastructure Limitations

The success of the application of AI-based Data Science depends on the integration capabilities of the existing hardware, software, and work process infrastructure. Usually most offline applications do not require significant changes in the existing infrastructure. A special case is users' addiction to Excel and their requirement to interact only within its environment. For real-time implementation of AI-based Data Science in the area of the IoT, however, a careful analysis of the software limitations and the maintenance infrastructure is a must.

Skillset Deficit

The business and academic world have understood that the key obstacle to materializing the potential of Data Science is the big deficit of data scientists. It is estimated that the number of graduates from Data Science programs could increase by a robust 7% per year. However, it is projected that even greater (12%) annual growth in demand, which would lead to a shortfall of some 250,000 data scientists. It has to be considered that in many courses this discipline is taught using examples far away from business reality, without messy data where nothing is obvious and one gap in data follows another. Another limitation of courses is a lack of business knowledge, so critical for the success of data scientists. The options for training of AI-based Data Scientists are very limited, since the AI-driven technologies covered in most courses are restricted to machine learning, with a focus on neural networks and decision trees.

Another missing skillset is that of the business translator, who serves as the link between data scientists and business users. In addition to being data-savvy, business translators need to have deep organizational knowledge and industrial or functional expertise. It may be possible to outsource analytics activities, but business translator roles require proprietary knowledge and should be more deeply embedded in the organization. It is estimated that there could be a demand for approximately two million to four million business translators in the United States alone over the next decade.[14]

Limited or No Modeling Experience

The success and the speed of implementing AI-based Data Science depend also on the previous record of modeling applications. Even lessons from applying simple statistical models are of help, since a modeling culture has been introduced. As a result, users have some experience in using and maintaining models, as well as assessment of the value created. On the other hand, one of the issues of limited modeling experience is the risk of unrealistic expectations, which paves the way for an application fiasco. Usually the lack of a modeling culture is combined with limited infrastructure for implementation and support, which additionally raises the total cost of ownership due to the necessary investment in infrastructure and training.

1.4.2 Nontechnical Issues in Applying AI-Based Data Science

The key nontechnical issues that may lead to unsuccessful AI-based Data Science applications are shown in Fig. 1.13 and discussed below.

[14] McKinsey Global Institute, *The Age of Analytics: Competing in Data-Driven World*, 2016.

No Management Support

As with any emerging technology, AI-based Data Science needs initial management blessing before it can begin to deliver sustainable value. Of critical importance is also the consistency of support for a period of at least three years. Unfortunately, this requirement may be unrealistic in businesses with frequent restructuring and management changes. The best strategy to address this issue is to find application areas with fast demonstration of value creation and to promote AI-based Data Science with effective marketing.

Wrong Expectations

Probably the most difficult issue in applied AI-based Data Science is how to help the final user in defining the proper expectations of the technology. Very often the dangerous combination of lack of knowledge, technology hype, and negative reception from some research and business communities creates incorrect anticipation about the real capabilities of AI-based Data Science. The two extremes of wrong expectations, either by exaggeration or by underestimation of the capabilities of AI-based Data Science, cause almost equal damage to the promotion of the technology in industry.

Magic Bullet

The expectation of technical magic is based on the unique capabilities of applied AI-based Data Science to handle uncertainty and complexity and to generate novelty. The impressive features of the broad diversity of methods, such as machine learning, deep learning, evolutionary computation, and swarm intelligence, contribute to such a Harry Potter-like image even when most of the users do not understand the principles behind them. Another factor adding to the silver bullet perception of AI-based Data Science is the technology hype from the vendors, the media, and some high-ranking managers.

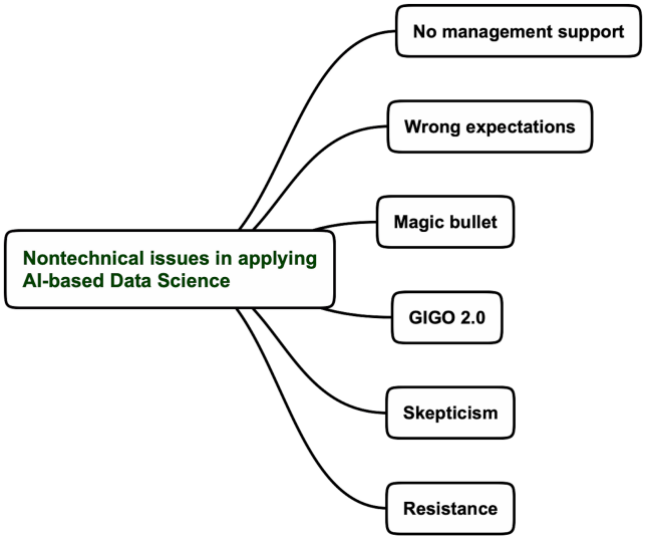


Fig. 1.13 Key nontechnical issues in applying AI-based Data Science

As a result, potential users approach AI-based Data Science as the last hope to resolve very complex and difficult problems. Often, they begin looking at the technology after several failed attempts of using other methods. In some cases, however, the problems are ill-defined and not supported by data and expertise. In order to avoid the magic bullet trap, it is strongly recommended to identify the requirements, communicate the limitations of the appropriate methods, and define realistic expectations in the very early phase of potential AI-based Data Science applications.

GIGO 2.0

The worst-case scenario of the magic bullet image is the GIGO 2.0 effect. In contrast to the classical meaning of GIGO 1.0 (Garbage-In-Garbage-Out), which represents ignorant expectations of a potential solution in the case of bad data, GIGO 2.0 embodies the next level of arrogant hope, defined as Garbage-In-Gold-Out. In essence, this is the false belief that low-quality data can be compensated for with sophisticated data analysis. Unfortunately, AI-based Data Science, with its diverse capabilities to analyze data, is one of the top-ranking technologies that create GIGO 2.0 arrogant expectations. It is observed that the bigger the disarray of the data, the higher the hope that exotic, unknown technologies will be able to clean up the mess. Usually top management who are unaware of the nasty reality of the mess initiate this behavior.

It is strongly recommended to protect potential AI-based Data Science applications from the negative consequences of the GIGO 2.0 effect. The best winning strategy is to define the requirements and the expectations in advance and to communicate clearly to the user the limitations of the methods. Better to reject an impossible implementation due to low-quality data than to poison the soil for many feasible AI-based Data Science applications in the future.

Skepticism

In contrast to the magic bullet optimistic euphoria, disbelief and lack of trust in the capabilities of applied AI-based Data Science is the other extreme of wrong expectation, but in this case on the negative side. Skepticism is usually the initial response of the final users of the technology on the business side. Several factors contribute to this behavior, such as lack of awareness of the technical capabilities and application potential of AI-based Data Science, lessons from other overhyped technology fiascos in the past, and caution about ambitious high-tech initiatives pushed by management.

Skepticism is a normal attitude if risk is not rewarded. Introducing emerging technologies, such as AI-based Data Science requires a risk-taking culture from all participants in this difficult process. The recommended strategy for success and reducing skepticism is to offer incentives to the developers and the users of the technology.

Resistance

The most difficult form of wrong expectations of AI-based Data Science is that technical or political biases leads people to actively oppose the technology. In most cases the driving forces of resistance are the parts of the business that feel threatened by the new, powerful capabilities of the technology. In order to prevent a future organizational fight, it is very important to communicate a firm commitment not to reduce the labor force as a result of AI-based Data Science applications.

The other part of the resistance movement against AI-based Data Science includes the professional statisticians. Most of them do not accept the statistical validity of some of the AI-based Data Science methods, especially neural-network-based models. The key argument of these fighters for the purity of the statistical theory is the lack of a statistically sound confidence metric for the nonlinear empirical solutions derived by AI-based Data Science.

The third part of the resistance camp against AI-based Data Science includes the active members of the Anything But Model (ABM) movement, who energetically oppose any attempt to introduce new technologies.

1.5 Common Mistakes

1.5.1 Believing the Hype

The biggest mistake that can be made at the beginning of applying AI-based Data Science is to trust blindly in the AI voodoo. There is no doubt that this paradigm has grown significantly in recent years and has begun aggressively to invade industry. However, the progress is driven mostly by the AI Olympians. One of the reasons for this concentration on the development of AI-driven technologies is that they have

increased the productivity of these high-tech giants and have opened up new markets (for example, for driverless vehicles). The hype and exuberant enthusiasm about AI is mostly based on the success of this paradigm in the AI Olympian companies.

The rest of the business world, however, has to look more carefully at the hype. The first step toward a reality correction is understanding the limitations of the key AI-driven technologies (details are given in Chap. 4.) The next step is gaining awareness about their maturity. For example, deep learning is still a technology in progress and not entirely ready for industry. The third step in replacing the hype with realistic expectations is to assess the requirements for nonstandard infrastructure and additional training.

Only after filling the gaps in one's knowledge about AI by these three steps can one have a more accurate technical view of this fashionable field.

1.5.2 Neglecting to Estimate the Demand for AI-Based Data Science

Another typical mistake is introducing AI-based Data Science without assessing the demand for its capabilities in a specific business. The fact that this package of technologies is successful at the AI Olympian companies does not automatically translate into productivity growth in your business. A comprehensive analysis of the potential need for AI-driven applications needs to be done in parallel with understanding the technical benefits of this paradigm. A good starting point is to look for use cases of similar AI-based Data Science business implementations. These can give more specific knowledge about the necessary AI-driven methods and infrastructure to start with.

1.5.3 Mass-Scale Introduction of AI-Based Data Science in a Business without Required Skillset Availability

When top management believes the AI hype and passionately embraces the paradigm, it often triggers a program of fast implementation of the technology across the organization. Unfortunately, this approach has a very low chance of success, due to the complex nature of AI-based technologies and the higher requirements for training. These technologies are not appropriate for accelerated mass-scale application. A combination of data scientists and business translators is needed to accomplish this task. Without gradually building an AI-based Data Science group with sufficient capacity in terms of of data scientists and business translators and the corresponding technical infrastructure, attempts to impose this technology from above by force are doomed to failure.

Recommendations for how to effectively introduce AI-based Data Science in a business are given in Chap. 16.

1.5.4 Introducing Data Science Bureaucracy

An inefficient way to introduce AI-based Data Science is by creating a bureaucratic structure and filling it with technically incompetent managers. A satirical scenario of this approach is shown in Fig 1.14.[15]

[15] This idea is inspired by a Tom Fishburne cartoon at www.marketoonist.com.

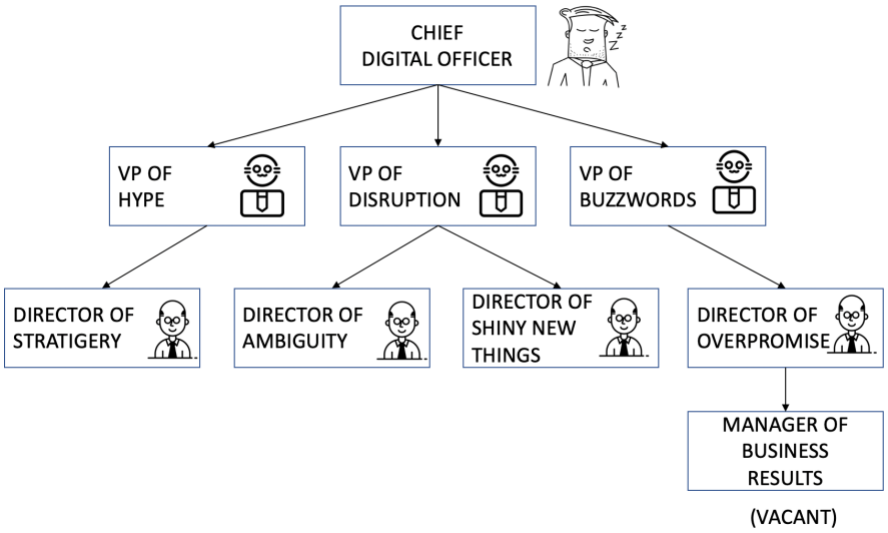


Fig. 1.14 The ideal bureaucratic structure of an AI-based Data Science business

Suggested Reading

B. Baesens, *Analytics in a Big Data World*, Wiley, 2014.
M. Berry and G. Linoff, *Data Mining Techniques*, 3rd edition, Wiley, 2013.
D. Hardoon and G. Shmueli, *Getting Started with Business Analytics*, CRC Press, 2013.
M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, 3rd edition, Pearson Education Canada, 2014.
F. Provost and T. Fawcett, *Data Science for Business*, O'Reilly Media, Inc, 2013.
S. Russell and P. Norvig, *Artificial Intelligence: a Modern Approach*, 3rd edition, Pearson, 2009.
A. Wodecki, *Artificial Intelligence in Value Creation: Improving Competitive Advantage*, Palgrave Macmillan, 2019.

Questions

Question 1

Give examples of hype related to Data Science, AI, big data, analytics, and the IoT.

Question 2

Find use cases that can benefit your business.

Question 3

What is the difference between descriptive and prescriptive analytics?

Question 4

Discuss the expected challenges in applying AI-based Data Science.

Question 5

AI Bingo. Bullshit Bingo is a very popular game in the corporate world. Usually one has a card with selected buzzwords related to a selected topic. The rule is to click on or mark each block when you see or hear the corresponding word or phrase. When you get five blocks horizontally, vertically, or diagonally, you stand up and shout "BULLSHIT!!!" An example of a card related to AI is given in Fig 1.15,.

Driverless cars	GPU	Machine Learning	Pattern	Data Mining
Paradigm	Watson	Data Science	NLP	Value
Analytics	Competitive Advantage	AI BINGO (free square)	The Cloud	Python
Cognitive Computing	Deep Learning	Robots	Framework	AI
Vision	Big Data	IoT	Data Scientist	Chatbots

Fig. 1.15 Example of AI bullshit bingo card

Create your own version of this card, following the steps in http://www.bullshitbingo.net/byo/.