What is Data Science?

Michael L. Brodie, Computer Science and Artificial Intelligence Laboratory, MIT

Draft: October 25, 2017

1 Introduction

Data Science, a new discovery paradigm, is potentially one of the most significant advances of the early 21st century. Originating in scientific discovery, it is being applied to every human endeavor for which there is adequate data. While remarkable successes have been achieved, even greater claims have been made. Risks and challenges abound. The science underlying *data science* has yet to emerge. Maturity is more than a decade away. This claim is based firstly on observing the centuries-long developments of its predecessor paradigms – empirical, theoretical, and Jim Gray's *Fourth Paradigm of Scientific Discovery* [19] (aka eScience, data-intensive, computational, procedural); and secondly on my studies of over 100 data science use cases, several data science-based startups, and, on my scientific advisory role for two Data Science Research Institutes (DSRIs) [5][17] that requires understanding the opportunities, state of the art, and research challenges for the emerging discipline of data science. Essential questions for a DSRI are: *What is data science*? and *What is world-class data science research*?

This chapter offers initial answers to these and related questions: What can data science do? What characteristics distinguish data science from previous scientific discovery paradigms? And What is the method for conducting data science? A companion chapter[11] addresses developing data science as a discipline, as a methodology, as well as data science research and education.

Data science has been used successfully to accelerate discovery of probabilistic outcomes in many domains. These outcomes have led to verified results through methods outside data science. Almost all data analyses are domain specific, many even specific to classes of models, classes of analytical methods, and specific pipelines. Few data science methods have been generalized outside the original domain, let alone to all domains. A rare and excellent exception is a generic scientific discovery method over scientific corpora [13] generalized from a specific method over medical corpora developed for drug discovery[16].

While there is much science in each domain-specific data science activity, there is little fundamental science that is applicable across domains. To warrant the designation *data science*, this emerging paradigm, as a science, requires fundamental principles and techniques applicable to all relevant domains. Since most data science work is domain specific, often model- and method-specific, data science does not yet warrant designation as a science.

This chapter explores the nature of data science, its qualitative differences with its predecessor scientific discovery paradigms, its core value and components, that when mature, would warrant the designation *data science*. Descriptions of large-scale data science activities in this chapter apply, scaled down, to data science activities of all sizes, including ubiquitous desktop data analyses.

2 What is data science?

Data science can be defined in terms of a data science method – a data-intensive extension of the scientific method.

> *Data Science is* a *body of principles and techniques for applying data analytics to accelerate the investigation of phenomena by acquiring new data, correcting and combining it with previous data, with measures of correctness, completeness, and efficiency of the derived results (correlations) with respect to some pre-defined (theoretical, deductive, top down) or emergent (inductive, bottom up) specification (scope, question, hypothesis, requirement).*

This definition is intended to discuss the nature of this remarkable new discovery paradigm. It benefits from two years research and experience over a previous version [12] and, no doubt, requires much refinement.

3 What can data science do, compared to the scientific method?

The data science method and its results are profoundly different to those of the scientific method. Scientific experiments analyze real phenomena directly under empirically defined controls. If statistical significance is achieved, the observed results, following the hypotheses, are accepted as casual - why the phenomena occurred. In contrast, data science does not directly analyze a phenomenon but data purported to represent features pertinent to the analysis of a phenomenon. Data can be empirical, i.e., observations acquired under known controls, but more often it is acquired with limited or unknown controls, and indirectly represents features of the phenomenon. Data Science can be used to discover correlations - what phenomena might occur or might have occurred – but, on its own, not to establish causality.

The prime benefit of data science is that it can accelerate discovery by processing potentially massive data volumes to find correlations amongst variables or patterns using methods that should also estimate the probability of the pattern being real or likely. This rapidly reduces the search space to a smaller set of likely patterns or correlations whose causality must be established by separate means. The Baylor-Watson study[16] discovered two potential cancer drugs in three months compared to one every two years. Verification of the potential cancer drugs is done through conventional methods, i.e., clinical trials.

Another difference and advantage concerns scale. While humans have difficulty reasoning over five to ten variables, some data science methods can "reason" over unlimited numbers of variables, finding correlations that humans and current methods could not. Readily available data, advanced algorithms, and modern hardware enable discovering correlations at orders of magnitude greater than ever before possible.

Data science analyses can be deductive or inductive. Deductive reasoning (aka empirical, top-down, theoretical) is used when specific hypotheses are to be evaluated. Inductive reasoning (bottom-up) is used not to evaluate specific hypotheses but using an analytical model and method to identify patterns or correlations that occur in the data with a frequency that meet some pre-defined specification, e.g., statistical significance. As opposed to evaluating pre-defined hypotheses in the top down approach, the bottom up approach is often said to "automatically" generate hypotheses, as in [16]. The inductive capacity of data science is often touted as its magic as the machine or methods such as machine learning, automatically and efficiently generate plausible hypotheses from data. While the acceleration and the scale of data being analyzed are major breakthroughs in discovery, the magic should be moderated by the fact that the generated hypotheses are derived from the models and methods used to generate them. The appearance of magic may derive from the fact that we do not understand how some analytical methods, e.g., machine learning, derive their results. This is a fundamental data science research challenge as we would like to understand the reasoning that led to a discovery, as is required in medicine, and in 2018 in the European Union, by law (the General Data Protection Regulation (GDPR)).

4 What are the components of data science?

Extending the analogy with science and the scientific method, data science, when mature, will be a systematic discipline with components that are applicable to most domains – most human endeavors. There are four categories of data science components, all incomplete awaiting research and development: data science principles, models, and methods; data science pipelines; data science infrastructure; and data infrastructure. Below, we discuss these components in terms of their support of a specific data science activity.

Successful data science activities have developed and deployed these components specific to the domain and study. To be considered a science, these components must be generalized across multiple domains, just as the scientific method applies not only to all sciences, but in recent decades has been applied to domains previously not considered sciences, e.g., economics, humanities, literature, psychology, sociology, and history.

A data science activity must be based on ***data science principles, models****, and (analytical)* ***methods***. Principles include those of science and of the scientific method applied to data science, for example, deductive reasoning, objectivity or lack of bias, reproducibility, and provenance. Particularly important are collaborative and cross-disciplinary methods. How do scientific principles apply to discovery over data? What principles underlie evidence-based reasoning for planning, predicting, decision-making, and policy-making in any domain?

A data science activity uses one or more models. A model represents the parameters that are the critical properties of the phenomenon to be analyzed. It often takes multiple models to capture all relevant features. Models are typically domain specific, often already established in the domain. Increasingly, models are developed for a data science activity, e.g., feature extraction from a data set is common for many AI methods. Data science activities often require the continuous refinement of a model to meet the analytical requirements of the activity. This leads to the need for model management to capture the settings and results of the planned and evaluated model variations. It is increasingly common, as in biology, to use multiple, distinct models, called an ensemble of models, each of which provides insights from a particular perspective. Each model, like each person in Plato's Allegory of the Cave, represents a different perspective of the same phenomenon, what Plato called shadows. Each model – each person – observes what appears to be the same phenomenon, yet each sees it differently. No one model – person – sees the entire thing, yet collectively they capture the whole phenomenon from many perspectives. It is rarely necessary, feasible, or of value to integrate different perspectives into a single integrated model. After all, there is no ultimate or truthful model save the phenomenon itself. Ensemble or shadow modelling is a natural and nuanced form of data integration[7], analogous to ensemble modelling in biology and ensemble learning and forecasting in other domains.

A data science activity can involve many analytical methods. A given method or algorithm is designed to analyze specific features of a data set. There are often variations of a method depending on the characteristics of the data set, e.g., sparse or dense, uniform or skewed, data type, data volume, etc., hence methods must be selected, or created, tuned for the data set and analytical requirements, and validated. In an analysis, there will be as many methods as there are specific features with corresponding specific data set types. Compared with analytical methods in science, their definition, selection, tuning, and validation in data science often involves scale in choice and computational requirements. Unless they are experts in the related methods, it is unlikely that a practicing data scientist understands the analytical method, e.g., machine learning, that they are applying relative to the analysis and data characteristics, let alone the thousands of available alternatives. This is a significant challenge in applying analytical sophisticated methods in business[14]. One of the greatest challenges in data science is the evaluation and interpretation of the likelihood of the outcome and the error bounds of the predictions that result from the analytical methods.

The central organizing principle and value proposition of a data science activity is its ***workflow*** or ***pipeline*** and its life cycle management [15]. A data science pipeline is an end-to-end sequence of steps from data discovery to the publication of the qualified, probabilistic interpretation of the result. A generic data science pipeline is comprehensive of all data science activities, hence can be used to define the *scope of data science*. The state of the art of data science is such that every data science activity has its own unique pipeline, as each data science activity is unique; however, there is far more variation than exists across pipelines in conventional science. Data science would benefit from a better understanding of pipelines and guidance on their design and development.

Data science pipelines often focus on the data analysis used to derive the results. However, over 80% of the resources required by a data science activity are consumed not by data analysis, but by the first two of the following five steps of a prototypical data science pipeline.

1. Raw data discovery, acquisition, and preparation into data repositories
2. Analytical data acquisition from data repositories
3. Data analysis
4. Results interpretation
5. Publish the results. Operationalize the pipeline for continuous analyses.

The core technical component for a data science activity is a ***data science infrastructure*** that supports the steps of the data science pipeline throughout its life cycle. A data science infrastructure consists of a workflow platform that supports the definition, refinement, execution, and reporting of data science activities in the pipeline. The workflow platform is supported by the infrastructure required to support workflow tasks such as data discovery, data mining, data preparation, data management, networking, libraries of analytical models and analytical methods, visualization, etc. To support user productivity, a user interface is required for each class of user, each with their own user experience. There are more than 60 such data science platforms - a new class of product - of which 16 meet Gartner's and Forrester's requirements [1][2][18]. These products are complex with over 15 component products such as database management, model management, machine learning, advanced analytics, data exploration, visualization, and data preparation.

Data, the world's most valuable resource[20], is also the most valuable resource for the data science activities of an organization (e.g., commercial, educational, research, governmental) and for entire communities. While new data is always required for an existing or new data science activity, data science activities of an organization require a ***data infrastructure*** – a sustainable, robust data infrastructure consisting of repositories of raw and curated data required to support the data requirements of the organization's data science activities with the associated support processes such as data stewardship. Many organizations are just developing data science data infrastructures. The best known are those that support large research communities. The US National Research Foundation is developing the *Sustainable Digital Data Preservation and Access Network Partners* to support data science for national science and engineering research and education. The 1000 Genomes Project Consortium created the world's largest catalog of genomic differences among humans, providing researchers worldwide with powerful clues to help them establish why some people are susceptible to various diseases. There are more than ten additional genomics data infrastructures, including the Cancer Genome Atlas of the US National Institutes of Health, Intel's Collaborative Cancer Cloud, and the Seven Bridges Cancer Cloud. Amazon hosts the 1000 Genome Project and 30 other public data infrastructures on topics such as Geospatial and Environmental Datasets, Genomics and Life Science Datasets, and Datasets for Machine Learning.

5 What is the method for conducting data science?

A data science activity is developed based on data science principles, models and analytical methods. The result of its design and development is a data science pipeline that will operate on a data science infrastructure, or platform, and will access data from some data infrastructure. There are a myriad of design and development methods to get from the principles to the pipeline. What follows is a description of a fairly generic data science method.

The ***data science method***, until better alternatives arise, is modelled on the scientific method. The following is one example of the applying the empirical approach to data science analysis, analogous to experimental design for science experiments. Each step requires verification, e.g., using experts, published literature, previous analysis; and continuous iterative improvement to reach results that meet a predefined specification. Each step may require revisiting a previous step, depending on its outcome. As with any scientific analysis, every attempt should be made to avoid bias, namely, attempting to prove preconceived ideas beyond the model, methods, and hypotheses. The method may run for hours to days for a small analysis; months, as for the Baylor-Watson drug discovery[16]; or years, as for the Kepler Space Telescope[8] and LIGO[9]. Design and development times can be similar to run times. Otto, a German e-commerce merchant, developed over months an AI-based system that predicts with 90% accuracy what products will be sold in the next 30 days and a companion system that automatically purchases over 200,000 items a month from third-party brands without human intervention. Otto selected, modified, and tuned a deep-learning algorithm originally designed for particle-physics experiments at CERN[3]. These systems run continuously.

A Generic Data Science Method

- Identify the phenomena or problem to be investigated. What is the desired outcome?
- Using domain knowledge, define the problem in terms of models that represent the critical factors or parameters to be analyzed (the WHAT of your analysis), based on the data likely to be available for the analysis. Understanding the domain precedes defining hypotheses to avoid bias.
- If the analysis is to be top down, formulate the hypotheses to be evaluated over the parameters and models.

- Design the analysis in terms of an end-to-end workflow or pipeline from the data discovery and acquisition, through analysis and results interpretation. The analysis should be designed to identify probabilistically significant correlations (What) and set requirements for acceptable likelihood and error bounds.
- Ensure the conceptual validity of the data analysis design.
- Design, test, and evaluate each step in the pipeline, selecting the relevant methods, i.e., class of relevant algorithms, in preparation for developing the following steps.
  - Discover, acquire, and prepare data required for the parameters and models ensuring that the results are consistent with previous steps.
  - For each analytical method, select and tune the relevant algorithm to meet the analytical requirements. This and the previous step are highly interrelated and often executed iteratively until the requirements are met with test or training data.
  - Ensure the validity the data analysis implementation.
- Execute the pipeline ensuring that requirements, e.g., probabilities and error bounds, are met.
- Ensure empirical (common sense) validation - the validity of the results with respect to the phenomena being investigated.
- Interpret the results with respect to the models, methods, and data analytic requirements. Evaluate the results (patterns or correlations) that meet the requirements for causality to be validated by methods outside data science.
- If the pipeline is to operate continuously, operationalize and monitor the analysis.

6 What is data science in practice?

Each data science activity develops its own unique data science method. Three very successful data science activities are described below in point form descriptions, using the above terminology to illustrate the components of data science in practice. They were conducted over 18, 20, and 2 years respectively. Their data science pipelines operated for 4 years, 3 years (to date), and 3 months respectively.

6.1 Kepler Space Telescope: Discovering Exoplanets

The Kepler space telescope project, initiated in 1999, and its successor project K2, have catalogued thousands of exoplanets by means of data analytics over Big Data. A detailed description of Kepler and access to its data is at NASA's Kepler & K2 Web site.

- **Objective and phenomenon**: Discover exoplanets in telescopic images
- **Project:** NASA-led collaboration of US government agencies, universities, and companies.
- **Critical parameters**: Over 100, e.g., planet luminosity, temperature, planet location relative to its sun.
- **Models**: Over 30 primarily established astrophysical models, e.g., the relationship between luminosity, hence, size, and temperature, a fundamental to stellar parameter, was established a century ago by Ejnar Hertzsprung and Henry Russell.
- **Methods**: Over 100, e.g., multi-scale Bayesian Maximum A Priori method used for systematic error removal from raw data. AI was not a principle method in this project.
- **Hypotheses**: Five, including "Determine the percentage of terrestrial and larger planets that are in or near the habitable zone of a wide variety of stars"
- **Data**: 100's of data types described in the Data Characteristics Handbook in the NASA Exoplanet Archive
- **Pipeline**: The **Kepler Science Pipeline** failed almost immediately after launch due to temperature and other unanticipated issues. After being repaired from earth, it worked well for 4 years.
- **Data discovery and acquisition**: Required approximately 90% of the total effort and resources
- **Algorithm selecting and tuning**: Models and methods were selected, developed, tuned and tested for the decade from project inception in 1999 to satellite launch in 2009, and were refined continuously.

- **Verification:** Every model and method was verified, e.g., exoplanet observations were verified using the Keck observatory in Hawaii.
- **Probabilistic outcomes**
  Kepler:
  - Candidates (<95%): 4,496
  - Confirmed (>99%): 2,330
  - Confirmed: <2X Earth-size in habitable zone: 30
  - Probably (<99%): 1,285
  - Probably not (~99%): 707

  K2:
  - Candidate (<95%): 521
  - Confirmed (>99%): 140

6.2 LIGO: Detecting Gravitational Waves

The LIGO project, detected cosmic gravitational waves predicted by Einstein's 1916 Theory of General Relativity for which it's originators were awarded the 2017 Nobel Prize. The project and its data are available at the LIGO Scientific Collaboration website.

- **Objective and phenomenon**: Observe cosmic gravitational waves.
- **Project**: Initiated in 1997 with 1,000 scientists in 100 institutes across 18 countries.
- **Equipment**: Laser Interferometer Gravitational-Wave Observatory (world's most sensitive detector)
- **Go Live:** September 2015 (after massive upgrade)
- **Data:** 100,000 channels of measurement of which one is for gravitational waves
- **Models**: At least one model per channel
- **Methods**: At least one data analysis method per data type being analyzed. AI was not used.
- **Challenges:** Equipment and pipeline (as is typical in data science activities)
- **Results:**
  - September 2015 (moments after reboot following massive upgrade), a gravitational wave, ripples in the fabric of space-time, was detected estimated to be the result of two black holes colliding 1.3BN light years from Earth.
  - Since then, four more gravitational waves were detected, one as this chapter went to press.
- **Collaboration**: The project depended on continuous collaboration between experimentalists who developed the equipment and theorists who defined what a signal from two black holes colliding would look like, let alone collaboration scientists, institutes, and countries.

6.3 Baylor-Watson: Cancer Drug Discovery

The Baylor-Watson drug discovery project [16] is a wonderful example of data-driven discovery and automatic hypothesis generation that discovered two novel kinases as potential sources for cancer drug development. These results that were determined to have a very high likelihood of success were developed in three months using IBM's Watson compared with the typical multi-year efforts that typically discover one candidate in two years.

- **Objective and phenomenon**: Discover kinases that regulate protein p53 to reduce or stem cancerous cell growth that have not yet been evaluated as a potential cancer drug.
- **Project**: Two years starting in 2012 between IBM Watson and the Baylor College of Medicine
- **Equipment**: Watson as a data science platform; PubMed a data infrastructure containing a corpus of 23m medical research articles.
- **Data:** 23M abstracts reduced to 240,00 papers on kinases reduced to 70,000 papers on kinases that regulate protein p53
- **Hypothesis**: Some of 500 kinases in the corpus regulate p53 and have not yet used for drugs

- **AI Models / methods:** "network analysis[13]" including textual analysis, graphical models of proteins and kinases, similarity analysis
- **Pipeline**: Explore, Interpret, and Analyze
    - **Explore:** scan abstracts to select kinase papers using text signatures
    - **Interpret:** extract kinase entities from papers and build connected graph of similarity amongst kinases
    - **Analyze**: diffuse annotations over kinases to rank order the best candidates for further experimentation
- **Data discovery and acquisition:** textual analysis of PubMed
- **Challenge:** designing, developing and tuning models and methods to scan abstracts for relevant papers; to construct a graphical model of the relevant relationships, to select kinases that regulate p53.
- **Execution:** 3 months
- **Results**: Two potential cancer drugs in 3 months versus 1 every 2 years (acceleration)
- **Validation**: The methods discovered 9 kinases of interest in corpus to 2003; 7 of 9 were empirically verified in the period 2004-2013. This raised the probability that the other two that had not yet been verified clinically, were highly likely candidates.
- **Causality**: Work is underway to develop drugs that use the kinases to regulate p53 to stem or reduce cancerous cell growth.
- **Collaboration**: The project involved collaboration between genetic researchers, oncologists, experts in AI and natural language understanding, and computer scientists.

7 How important is collaboration in data science?

Data Science is an inherently collaborative activity involving multiple disciplines including domain knowledge to design, develop, validate, and execute data analytics pipelines. Data science activities almost always require expertise from multiple disciplines – subject domain, statistics, computing, AI, analytics, mathematics, and many more. Conventional scientific activities have acquired the few required disciplines, e.g., a biology laboratory hires the required data management skills. This has been a highly appropriate way to evolve, however, as data science matures, so should collaboration amongst disciplines, domains, and individuals beyond the current top-down, domain specific form. Data science activities within a specific domain tend not to have, and have difficulty acquiring, anticipated and unanticipated expertise such as machine learning, statistics at scale, and Big Data management. The popularity of data science has led to courses in every university, yet the required skills and experience are rare, and like data science itself, just emerging as topics let alone as university and training courses.

The need for collaboration on basic research and engineering on the fundamental building blocks of data science and data science infrastructures can be seen in a recent report from University of California, Berkeley[4] researchers from many domains – statistics, AI, data management, systems, security, data centers, distributed computing, and more.

Data science activities have emerged in most research labs in most universities. Until 2017, many Harvard University departments had one or more groups conducting data science research and offered a myriad of data science degrees and certificates. In March 2017, the Harvard Data Science initiative was established to coordinate the many activities. This pattern has repeated at over 100 major universities worldwide, resulting in over 100 Data Science Research Institutes being established since 2015 – themselves just emerging. The creation of over 100 DSRI in approximately two years, all heavily funded by governments and by partner industrial organizations, is an indication of the belief in the potential of data science not just as a new discovery paradigm, but as a basis for business and economic growth.

Collaboration is an emerging challenge in data science not only at the scientific level but also at the strategic and organizational levels. Analysts report that most early industry big data deployments failed due to a lack of domain-business-analytics-IT collaboration[14]. Most of the over 100 Data Science Research Institutes (DSRIs) involve a grouping of departments or groups with an interest in data science in their domain, into a higher level DSRI. In principle, the DSRI would strive for higher-level, scientific and strategic goals, such as contributing to data science (i.e., the science underlying data science) in contrast with the contributions made in a specific domain by each partner organization. But how does the DSRI operate? How should it be organized so as to encourage collaboration and achieving higher-level goals.

While data science is inherently multi-disciplinary, hence collaborative, in nature, scientists and practitioners lack training in collaboration and are motivated to focus on their objectives and domain. Why would a bioinformaticist (bioinformatician) attempt to establish a data science method that goes beyond her requirements, especially as it requires a deep understanding of domains such as deep learning? Collaboration is also a significant organizational challenge specifically for the over 100 DSRIs that were formed as a federation of organizational units each of which conduct data science activities in different domains. Like the bioinformaticist, each organization has its own objectives, budget, and investments in funding and intellectual property. In such an environment, how does a DSRI establish strategic directions and set research objectives? Through a DSRI Chief Scientific Officer [11]?

8 What is world-class data science research?

While many data science groups share a passion for data science, they do not share common data science components – principles, models, methods; pipelines; data science infrastructures, data infrastructures, or data science methods. This is perfectly reasonable given the state of data science, and the research needs of the individual groups; however, to what extent are these separate groups pursing data science, *per se*? This raises our original questions: What is data science? and What is world-class data science research? These questions are central to planning and directing data science research such as in DSRIs.

There are two types of data science research, domain specific contributions and contributions to the discipline of data science itself. Domain specific, world class data science research concerns applications of data science in specific domains resulting in domain-specific discoveries that are recognized in its domain as being world class, of which there are many compelling examples [8][9][16]. To be considered data science, the research must be based on some version of the data science method and utilize the components of data science. The data science components or the data science method should be critical to achieving the result in comparison with other methods, e.g., accelerating discovering, finding solutions that might not have been discovered otherwise.

Equally or even more important world class data science research is to establish data science as a science or as a discipline with robust principles, models, and methods; pipelines; a data science method supported by robust data science infrastructures, and data infrastructures applicable to multiple domains. A wonderful example of generalizing a domain-specific data science method is extending the network analysis method applied to some specific medical corpora used successfully in drug discovery [16] to domain-independent scientific discovery applied to arbitrary scientific corpora [13].

The charter of every DSRI should include both domain-specific data science research and research to establish data science as a discipline. Since most DSRIs were formed from groups successfully practicing domain-specific data science, they are all striving for world class domain-specific data science. Without world class research in data science per se, it would be hard to argue that the DSRI contributes more than the sum of the parts. One might argue that lacking research into data science per se means that the DSRI has more of an organizational or marketing purpose than a research focus. Both research objectives were the intent of the government of Ireland and its funding agency, Science Foundation Ireland, in establishing, in 2012, the national DSRI, Insight Center for Data Analytics[5]. Achieving these objectives is challenging.

9 Conclusion

Data science is an emerging paradigm with the primary advantage of accelerating discovery of correlations between variables at a scale and speed beyond human cognition and previous discovery paradigms. Data science differs paradigmatically from its predecessor scientific discovery paradigms that were designed to discover in real contexts, causality – Why a phenomenon occurred. Data science is designed to discover in data, correlations – What phenomena may have or may occur. Unlike previous scientific discovery paradigms that were designed for scientific discovery and are now applied in many non-scientific domains; data science is applicable to any domain for which adequate data is available.

Hence, the potential of broad applicability and accelerating discovery in any domain to rapidly reduce the search space for solutions holds remarkable potential for all fields. While already applicable and applied successfully in many domains, there are many challenges that must be addresses over the next decade as data science matures.

My decade-long experience in data science, suggests that there are no compelling answers to the questions posed in this chapter. This is due in part to its recent emergence, it's almost unlimited breadth of applicability, and to its inherently multi-disciplinary, collaborative nature.

To warrant the designation *data science*, this emerging paradigm, as a science, requires fundamental principles and techniques applicable to all relevant domains. Since most "data science" work is domain specific, often model- and method-specific, "data science" does not yet warrant the designation of a science. This is not a mere appeal for formalism. There are many challenges facing data science such as validating results thereby minimizing the risks of failures. The purported benefits of data science, e.g., in accelerating the discovery of cancer cures and solutions to global warming, warrant establishing rigorous, efficient data science principles and methods.

10 References

[1] Critical Capabilities for Data Science Platforms, Gartner, June 7, 2017 ID: G00326671

[2] Gartner 2017 Magic Quadrant for Data Science Platforms, 14 February 2017 **ID:** G00301536

[3] How Germany's Otto uses artificial intelligence, The Economist, April 12, 2017.

[4] I. Stoica, D. Song, R. Ada Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. Gonzalez, K. Goldberg, A. Ghodsi, D. Culler, P. Abbeel, A Berkeley View of Systems Challenges for AI, Technical Report No. UCB/EECS-2017-159, October 16, 2017

[5] Insight Center for Data Analytics, Ireland.

[6] J. Duggan and M. L. Brodie, "Hephaestus: Data Reuse for Accelerating Scientific Discovery," CIDR 2015, Jan. 2015.

[7] J. T. Liu, "Shadow Theory, data model design for data integration," CoRR, vol. 1209, 2012. arXiv:1209.2647

[8] Kepler Space Telescope.

[9] LIGO - Laser Interferometer Gravitational-Wave Observatory

[10] M. Braschler, T. Stadelmann, K. Stockinger (Eds.), "Applied Data Science - Lessons Learned for the Data-Driven Business", Berlin, Heidelberg: Springer, expected 2018

[11] M. L. Brodie, *Necessity is the Mother of Invention*: On Developing Data Science, to appear in [10]

[12] M. L. Brodie**,** *Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery*, in Shannon Cutt (ed.), Getting Data Right: Tackling the Challenges of Big Data Volume and Variety, O'Reilly Media, Sebastopol, CA, USA, June 2015

[13] M. Nagarajan, et al. 2015. Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15). ACM, New York, NY, USA, 2019-2028.

[14] Predictions 2016: The Path from Data to Action for Marketers: *How Marketers Will Elevate Systems of Insight.* Forrester Research, November 9, 2015

[15] Realizing the Potential of Data Science, Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group, December 2016

[16] S. Spangler, et al. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '14). ACM, New York, NY, USA, 1877-1886.

[17] Swinburne Data Science Research Institute, Melbourne, Australia.

[18] The Forrester Wave^{TM}: Predictive Analytics and Machine Learning Solutions, Q1 2017, March 7, 2017

[19] The Fourth Paradigm: Data-Intensive Scientific Discovery Edited by Tony Hey, Stewart Tansley, and Kristin Tolle, Microsoft Research, 2009

[20] The World's most valuable resource, The Economist, May 4, 2017.