Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery

© Michael L. Brodie
CSAIL, MIT
Cambridge, MA, USA
mlbrodie@csail.mit.edu

Abstract

Over the past two decades, Data-Intensive Analysis has emerged not only as a basis for the *Fourth Paradigm* of engineering and scientific discovery but as a basis for discovery in most human endeavors for which data is available. Originating in the 1960s, its recent emergence due to Big Data and massive computing power is leading to widespread deployment, yet it is in its infancy in its application and our understanding of it; hence in its development. Given the potential risks and rewards of Data-Intensive Analysis and its breadth of application, it is imperative that we get this right.

The objective of this emerging Fourth Paradigm is more than acquiring data and extracting knowledge. Like its predecessor the scientific method, the objective of the Fourth Paradigm is to *investigate phenomena by acquiring new knowledge, and correct and integrate it with previous knowledge*. In addition, data science is a body of *principles and techniques with which to measure and improve the correctness, completeness, and efficiency of Data-Intensive Analysis.* It is now time to identify and understand the fundamentals. In my research, I have analyzed more than 30 very large-scale use cases to understand current practical aspects, to gain insight into the fundamentals, and to address the fourth "V" of Big Data – veracity -- the accuracy of the data and the resulting analytics This development may take decades.

1 Data Science: A New Discovery Paradigm That Will Transform Our World

1.1 Introduction

Over the past two decades, Data-Intensive Analysis (also called Big Data Analytics) has emerged not only as a basis for the *Fourth Paradigm [8]* of engineering and scientific discovery but more broadly as a basis for discovery in most human endeavours for which data is available. Roots of Data-Intensive Analysis (DIA) that have led to its recent dramatic growth include Big Data (c. 2000) that, just emerging, is opening the door to profound change – to new ways of reasoning, problem solving, and processing that in turn bring new opportunities and challenges.

To better understand DIA and its opportunities and challenges I examined over 30 DIA use cases that are at very large-scale - in the range where theory and practice may break. This paper summarizes some key results of my research related to understanding and defining Data Science as *a body of principles and techniques with which to measure and improve the correctness, completeness, and efficiency of Data-Intensive Analysis.* As with its predecessor discovery paradigms, establishing this emerging Fourth Paradigm and the underlying principles and techniques of Data Science may take decades.

1.2 Significance of DIA and Data Science

Data Science is transforming discovery in many human endeavours including healthcare, manufacturing, education, financial modelling, policing, and marketing [10][13]. It has been used to produce significant results in areas from particle physics (e.g., Higgs Boson), to identifying and resolving sleep disorders using Fitbit data, to recommenders for literature, theatre, and shopping. More than 50 national governments have established data-driven strategies as an official policy as in science and engineering [2] as well as in healthcare, e.g., US National Institutes of Health and President Obama's Precision Medicine Initiative [15] for "Delivering the right treatments, at the right time, every time to the right person." The hope, supported by early results, is that data-driven techniques will accelerate the discovery of treatments to manage and prevent chronic diseases with more precision and that are tailored to specific individuals as well as being at dramatically lower cost.

Data Science is being used to radically transform entire domains, such as medicine and biomedical research as stated as the purpose of the newly created Center for Biomedical Informatics at the Harvard Medical School. It is also making an impact in economics [14], drug discovery [17], and many other domains. As a result of its successes and potential Data Science is rapidly becoming a sub-discipline of most academic areas. These developments suggest the strong belief in the potential value of Data Science – but can it deliver?

Early successes and clearly stated expectations of Data Science are truly remarkable; however, its actual deployment, like many hot trends, is far less than it appears. According to Gartner's 2015 survey of Big

Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015

Data Management and Analytics, 60% of the Fortune 500 claim to have deployed Data Science, less than 20% have implemented consequent significant changes and less than 1% have optimized its benefits. Gartner concludes that 85% will be unable to exploit Big Data in 2015. The vast majority of deployments address tactical aspects of existing processes and static business intelligence rather than realizing its power by identifying strategic advantages through discovering previously unforeseen value.

1.3 Illustrious Histories: The Origins of Data Science

Data Science is in its infancy. Few individuals or organizations understand the potential of and the paradigm shift associated with Data Science, let alone understand it conceptually. The high rewards and the equally high risks and its pervasive application make it imperative that we better understand Data Science – its models, methods, processes, and results.

Data Science is inherently multi-disciplinary drawing on over 30 allied disciplines, according to some definitions. Its principle components include mathematics, statistics, and computer science especially areas such as AI (e.g., machine learning), data management, and high performance computing. While these disciplines need to be evaluated in the new paradigm, they have long illustrious histories. Data analysis developed over 4,000 years ago with origins in Babylon (17^{th}-12^{th} C BCE) and India (12^{th} C BCE). Mathematical analysis originated in the 17^{th} C around the time of the Scientific Revolution. While statistics has its roots in 5^{th} C BCE and the 18^{th} C, its application in Data Science originated in 1962 with John W. Tukey [20] and George Box[4]. These long illustrious histories suggest that Data Science draws on well-established results that took decades or centuries to develop. To what extent do they (e.g., statistical significance) apply in this paradigmatically new context?

Data Science constitutes a new paradigm in the sense of Kuhn's scientific revolutions [12]. Data Science's predecessor paradigm, the Scientific Method, has approximately 2,000 years in the development of empiricism starting with Aristotle (384-322 BCE), Ptolemy (1^{st} C), and the Bacons (13^{th}, 16^{th} C). Data Science, a primary basis of eScience [8], collectively termed the Fourth Paradigm, is emerging following the ~1,000-year development of its three predecessor paradigms of scientific and engineering discovery: theory, experimentation, and simulation [8]. Data Science that has developed and been applied for over 50 years qualitatively changed in the late 20^{th} century with the emergence of Big Data, typically defined as data at *volumes*, *velocities*, and *variety* that current technologies, let alone humans, cannot handle efficiently. This paper addresses another characteristic that current technologies and theories do not handle well, *veracity*.

1.4 What Could Possibly Go Wrong?

Do we understand the risks of recommending the wrong film, the wrong product, the wrong medical diagnoses, treatments, or drugs? The minimal apparent risk of a result that fails to achieve its objectives when acted upon includes losses in time, resources, customer satisfaction, customers, and potentially a loss of business. The vast majority of Data Science applications face such small risks; hence veracity has received little attention. Far greater risks could be incurred if incorrect Data Science results are acted upon in critical contexts, such as those already underway in drug discovery [18] and personalized medicine. Most scientists in these contexts are well aware of the risks of errors, hence go to extremes to estimate and minimize them. The wonder of CERN's ATLAS and CMS projects "discovery" of the Higgs Boson announced July 4, 2012 with a confidence of 5 sigma might suggest that the results were achieved overnight. They were not. They took 40 years and included Data Science techniques developed over a decade applied over Big Data by two independent projects, ATLAS and CMS, each of which were subsequently peer reviewed and published [1][11] with a further yearlong verification that established a confidence of 10 sigma. To what extent do the vast majority of Data Science applications concern themselves with verification and error bounds let alone understand the verification methods applied at CERN? Informal surveys of data scientists conducted in this study at Data Science conferences suggest that 80% of customers never ask for error bounds.

The existential risks of applying Data Science have been raised by world leading authorities such as the Organization for Economic Cooperation and Development, the AI [3][7][9][19] and legal [5] communities with the most extreme concerns stated by the Future of Life Institute with the objective of *safeguarding life and developing optimistic visions of the future* in order to *mitigate existential risks facing humanity* from AI.

Given the potential risks and rewards of DIA and of its breadth of application across conventional, empirical scientific and engineering domains as well as across most human endeavors we better get this right! The scientific and engineering communities place high confidence in their existing discovery paradigms with well-defined measures of likelihood and confidence within relatively precise error estimates[1]. Can we say the same for modern Data Science as a discovery paradigm and for its results? A simple observation of the formal development of the processes and methods of its predecessors suggest that we cannot. Indeed, we do not know if or under what conditions the constituent disciplines, like statistics, may break down.

Do we understand DIA to the extent that we can assign probabilistic measures of likelihood to its results? With the scale and emerging nature of DIA-

[1] Even after 1,000 years serious issues persist, e.g., P values (significance) and reproducibility.

based discovery, how do we estimate the correctness and completeness of analytical results relative to a hypothesized discovery question when the underlying principles and techniques may not apply in this new context?

In summary, we do not yet understand DIA adequately to quantify the probability or likelihood that a projected outcome will occur within estimated error bounds. While CERN used Data Science and Big Data to identify results, verification was ultimately empirical, as it must be in drug discovery [18] and other critical areas, until analytical techniques are developed and proven robust.

1.5 Do We Understand Data Science?

Do we even understand what Data Science methods compute or how they work? Human thought is limited by the human mind. According to Miller's Law [14], the human mind (short term working memory) is capable of conceiving less than ten (7 +/- 2) concepts at one time. Hence, humans have difficulty understanding complex models involving more than ten variables. The conventional process is to imagine a small number of variables[2] then abstract or encapsulate that knowledge into a model that can subsequently augmented with more variables. Thus most scientific theories develop slowly over time into complex models. For example, Newton's model of particle physics was extended for 350 years through Bohr, Heisenberg, Einstein, and others, up to Glashow, Salam, and Weinberg, to form The Standard Model of Particle Physics. Scientific discovery in particle physics is wonderful and has taken over 350 years. Due to its complexity no physicist has understood the entire Standard Model for decades, rather it is represented in complex, computational models.

When humans analyse a problem, they do so with models with a limited number of variables. As the number of variables increase, it is increasingly difficult to understand the model and the potential combinations and correlations. Hence, humans limit their models and analyses to those that they can comprehend. These human-scale models are typically theory-driven thus limiting their scale (number of variables) to what can be conceived.

What if the phenomenon is arbitrarily complex or beyond immediate human conception? I suspect that this is addressed iteratively with one model (theory) becoming abstracted as the base for another more complex theory, and so on (standing on the shoulders of those who have gone before), e.g., the development of quantum physics from elementary particles. That is, once the human mind understands a model, it can form the basis of a more complex model. This development under the scientific method scales at a rate limited by human conception thus limiting the number of variables and complexity. This is error-prone since phenomena may not manifest at a certain level of complexity hence models correct at one scale may be wrong at a larger scale or *vice versa*, a model wrong at one scale (hence discarded) may become correct at a higher scale (more complex model).

Machine learning algorithms can identify correlations between thousands, millions, or even billions of variables. This suggests that it is difficult to impossible for humans to understand what or how these algorithms discover. Imagine trying to understand such a model that results from selecting some subset of the correlations on the assumption that they may be causal thus constitute a model of the phenomenon with high confidence of being correct with respect to some hypotheses, with or without error bars.

1.6 Cornerstone of A New Discovery Paradigm

The Fourth Paradigm - eScience supported by Data Science - is paradigmatically different from its predecessor discovery paradigms. It provides revolutionary new ways [12] of thinking, reasoning and processing - new modes of inquiry, problem solving, and decision-making. It is not the Third Paradigm augmented by Big Data, but something profoundly different. Losing sight of this difference forfeits its power and benefits and loses the perspective that it is *A Revolution That Will Transform How We Live, Work, and Think* [13].

Paradigm shifts are difficult to notice as they emerge, just as the proverbial frog does not notice that its hot bath is becoming lethal. There are several ways to describe the shift. There is a shift of resources from (empirically) discovering causality (*Why the phenomenon occurs*) – the heart of the Scientific Method – to discovering interesting correlations (*What might have occurred*). This shift involves moving from a strategic perspective driven by human generated hypotheses (theory-driven, top-down) to a tactical perspective driven by observations (data-driven, bottom-up).

Seen at their extremes, the Scientific Method involves testing hypotheses (theories) posed by scientists while Data Science can be used to generate hypotheses to be tested based on significant correlations amongst variables that are identified algorithmically in the data. In principle, vast amounts of data and computing power can be used to accelerate discovery simply by outpacing human thinking in both power and complexity. The power of Data Science is growing rapidly due to the development of ever more powerful computing resources and algorithms, such as deep learning. So rather than optimize an existing process, Data Science can be used to identify patterns that suggest unforeseen solutions, thus automating serendipity as it is called when a human observes an anomaly that stimulated a bright idea to resolve it.

However, even more compelling is one step beyond the simple version of this shift, namely a symbiosis of the both paradigms. For example, Data Science can be used to offer highly probable hypotheses or correlations from which we select those with acceptable error

[2] Physical science PhDs typically involve < 5 variables.

estimates and that are worthy of subsequent empirical analysis. In turn, empiricism is used to pursue these hypotheses until some converge and some diverge at which point Data Science can be applied to refine or confirm the converging hypotheses, having discarded the divergent hypotheses, and the cycle starts again. Ideally, one would optimize the combination of theory-driven empirical analysis with data-driven analysis to accelerate discovery faster than either on their own.

While Data Science is a cornerstone of a new discovery paradigm, it may be conceptually and methodologically more challenging than its predecessors since it involves everything included in its predecessor paradigms – modelling, methods, processes, measures of correctness, completeness, and efficiency – in a much more complex context, namely that of Big Data. Following well-established developments, we should try to find the fundamentals of Data Science – its principles and techniques – to help manage the complexity and guide its understanding and application.

2 Data Science: A Perspective

Since Data Science is in its infancy and is inherently multi-disciplinary, there are naturally many definitions of Data Science that should emerge and evolve with the discipline. As definitions serve many purposes, it is reasonable to have multiple definitions each serving different purposes. Most Data Science definitions attempt to define *Why* (it's purpose), *What* (constituent disciplines), and *How* (constituent actions of discovery workflows).

A common definition of Data Science is *the activity of extracting knowledge from data*[3]. While simple, it does not convey the larger goal of Data Science or its consequent challenges. A DIA activity is far more than a collection of actions or the mechanical processes of acquiring and analyzing data. Like its predecessor paradigm, the Scientific Method, the purpose of Data Science and a DIA activity is to *investigate phenomena by acquiring new knowledge, and correcting and integrating it with previous knowledge* – continually evolving our current understanding of the phenomena based on newly available data. We seldom start from scratch, clearly the simplest case here. Hence, discovering, understanding, and integrating data must precede extracting knowledge all at massive scale, i.e., largely by automated means.

The Scientific Method that underlies the Third Paradigm is a body of principles and techniques that provide the formal and practical bases of scientific and engineering discovery. The principles and techniques have been developed over hundreds of years originating with Plato and are still evolving today with significant unresolved issues such as statistical significance, (i.e., P values) and reproducibility.

While Data Science had its origins 50 years ago with Tukey [19] and Box [4] it started to change qualitatively less than two decades ago with the emergence of Big Data and the consequent paradigm shift described above. The focus of this research into *modern* Data Science is on veracity – the ability to estimate the correctness, completeness, and efficiency of an end-to-end DIA activity and of its results. Hence, I use the following definition that is in the spirit of [17].

> **Data Science** is a *body of principles and techniques for applying data-intensive analysis to investigate phenomena, acquire new knowledge, and correct and integrate previous knowledge with measures of correctness, completeness, and efficiency of the derived results with respect to some pre-defined (top down) or emergent (bottom up) specification (scope, question, hypothesis).*

3 Understanding Data Science From Practice

3.1 Methodology to Better Understand DIA

Driven by a passion for understanding Data Science in practice, my year-long and on-going research study has investigated over 30 very large scale Big Data applications most of which have produced or are daily producing significant value. The use cases include particle physics; astrophysics and satellite imagery; oceanography; economics; information services; several life sciences applications in pharmaceuticals, drug discovery, and genetics; and various areas of medicine including precision medicine, hospital studies, clinical trials, intensive care unit and emergency room medicine.

The focus is to investigate relatively well-understood, successful use cases where correctness is critical and the Big Data context is at massive scale; such use cases constitute less than 5% of all deployed Big Data analytics. The focus was on these use cases, as we do not know where errors may arise outside normal scientific and analytical errors. There is a greater likelihood that established disciplines, e.g., statistics and data management, might break at very large scale where errors due to failed fundamentals may be more obvious.

The breadth and depth of the use cases revealed strong, significant emerging trends, some of which are listed below. These confirmed for some use case owners, and suggested to others, solutions and directions that they were pursuing but could not have seen without the perspective of 30+ use cases.

3.2 DIA Processes

A Data-Intensive-Activity is an analytical process that consists of applying sophisticated analytical methods to large data sets that are stored under some analytical models. While this is the typical view of Data Science projects or DIA use cases, this analytical component of the DIA activity constitutes ~20% of an end-to-end DIA pipeline or workflow. Currently it consumes ~20% of the resources required to complete a DIA analysis.

[3] Wikipedia.com

An end-to-end DIA activity involves two data management processes that precede the DIA process, namely Raw Data Acquisition and Curation, and Analytical Data Acquisition. Raw Data Acquisition and Curation starts with discovering and understanding data in data sources and ends with integrating and storing curated data in a repository that represents entities in the domain of interest and metadata about those entities with which to make a specific interpretations and that is shared by a community of users. Analytical Data Acquisition starts with discovering and understanding data within the shared repository and ends with storing the resulting information, specific entities and interpretations, into an analytical model to be used by the subsequent DIA process.

Sophisticated algorithms such as machine learning largely automate DIA processes, as they have to be automated to process such large volumes of data using complex algorithms. Currently, Raw Data Acquisition and Curation, and Analytical Data Acquisition processes are far less automated typically requiring 80% or more of the total resources to complete.

This understanding leads to the following definitions.

> **Data-Intensive Discovery (DID)** *is the activity of using Big Data to investigate phenomena, to acquire new knowledge, and to correct and integrate previous knowledge.*

"-Intensive" is added when the data is "at scale". Theory-driven DID is the investigation of human generated scientific, engineering, or other hypotheses over Big Data. Data-Driven DID employs automatic hypothesis generation.

> **Data-Intensive Analysis** *is the process of analyzing Big Data with analytical methods and models.*

DID goes beyond the Third paradigm of scientific or engineering discovery by investigating scientific or engineering hypotheses using DIA. A DIA activity is an experiment over data thus requiring all aspects of a scientific experiment, e.g., experimental design, expressed over data, a.k.a. **data-based empiricism**.

> **A DIA Process** (workflow or pipeline) *is a sequence of operations that constitute an end-to-end DIA activity from the source data to the quantified, qualified result.*

Currently, ~80% of the effort and resources required for the entire DIA activity are due to the two data management processes – areas where scientists / analysts are not experts. Emerging technology, such as for data curation at scale, aims to flip that ratio from 80:20 to 20:80 so as to *let scientists do science*; *analysts do analysis; etc*. This requires an understanding of the data management processes and their correctness, completeness, and efficiency in addition to those of the DIA process. Another obvious consequence is that proportionally 80% of the errors that could arise in DIA may arise in the data management processes, prior to DIA even starting.

3.3 Characteristics of Large-Scale DIA Use Cases

The focus of my research is successful, very large scale, multi-year projects with many with 100s to 1,000s, of ongoing DIA activities. These activities are supported by a **DIA ecosystem** consisting of a community of users (e.g., over 5,000 scientists in the ATLAS and CMS projects at CERN and similar numbers of scientists using the worldwide Cancer Genome Atlas) and technology (e.g., *science gateways*[4], collectively referred to in some branches of science as *networked science*). Some significant trends that have emerged from the analysis of these use cases are listed, briefly, below.

The typical view of Data Science appears to be based on the vast majority (~95%) of DIA use cases. While they share some characteristics with those in this study, there are fundamental differences such as the concern for and due diligence associated with veracity as mentioned above.

Based on this study data analysis appears to fall into three classes. **Conventional data analysis** over "small data" accounts for at least 95% of all data analysis, often using Microsoft Excel. DIA over Big Data has two sub-classes, **simple DIA**, i.e., the vast majority of DIA use cases mentioned above, and **complex DIA** such as the use cases analyzed in this study that are characterized by complex analytical models (e.g., sub-models of the Standard Model of Physics, economic models, an organizational model for enterprises worldwide, and models for genetics and epigenetics) and a corresponding plethora of analytical methods (e.g., the vast method libraries in CERN's Root framework). The complexity of the models and methods are as complex as the phenomena being analyzed.

The most widely used DIA tools for simple cases claim to support analyst self-service in point-and-click environments, some claiming "point us at the data and we will find the patterns of interest for you". This characteristic is infeasible in the use cases analyzed. A requirement common to the use cases analyzed is not only the principle of being machine driven and human guided, i.e., a **man-machine symbiosis**, but extensive attempts to optimize this symbiosis for scale, cost, and precision (too much human-in-the-loop leads to errors, too little leads to nonsense).

DIA ecosystems are inherently **multi-disciplinary** (ideally interdisciplinary), **collaborative**, and **iterative**. Not only does DIA (Big Data Analytics) require multiple disciplines, e.g., genetics, statistics and machine learning, so too do the data management processes require multiple disciplines, e.g., data management, domain and machine learning experts for

[4] There are over 60 large-scale scientific gateways, e.g., The Cancer Genome Atlas and CERN's Worldwide LHC Computing Grid.

data curation, statisticians for sampling, etc.

In large-scale DIA ecosystems, a DIA is a **virtual experiment** [6]. Far from claims of simplicity and point-and-click self-service, most large-scale DIA activities reflect the complexity of the analysis at hand and are the result of long-term (months to years) experimental designs that involve greater complexity than their empirical counterparts to deal with scale, significance, hypotheses, null hypotheses, and deeper challenges such as determining causality from correlations and identifying and dealing with biases and often irrational human intervention.

Finally, **veracity** is one of the most significant challenges and critical requirements of all DIA ecosystems studied. While there are many, complex methods in conventional Data Science to estimate veracity most owners of use cases studied expressed concern for adequately estimating veracity in modern Data Science. Most assume that all data is imprecise; hence require **probabilistic measures** and **error bars** and **likelihood** estimates for all results. More basically, most DIA ecosystem experts recognize that errors can arise across an end-to-end DIA activity and are investing substantially in addressing these issues in both the DIA processes and the data management processes that currently require significant human guidance.

An objective of this research is to discover the extent to which the above characteristics of very large scale, complex DIAs also apply to simple DIAs. There is a strong likelihood that they apply directly but are difficult to detect. That is the principles and techniques of DIA apply equally to simple and complex DIA.

3.4 Looking Into A Use Case

Due to the detail involved, there is not space in this chapter or book to describe a single use case considered in this study. However, let's look into a single step of a use case involving a virtual experiment conducted at CERN in the Atlas project. The heart of empirical science is experimental design. It starts by identifying, formulating, and verifying a worthy hypothesis to pursue. This first complex step typically involves a multi-disciplinary team, called the collaborators for this virtual experiment, often from around the world for more than a year. We consider the second step, the construction of the control or background model (executable software and data) that creates the background (e.g., executable or testable model and a given data set) required as the basis within which to search (analyze) for "signals" that would represent the phenomenon being investigated in the hypothesis. This is the control that completely excludes the data of interest. The data of interest (the signal region) is "blinded" completely so as not to bias the experiment. The background (control) is designed using software that simulates relevant parts of the standard model of particle physics plus data from Atlas selected with the appropriate signatures with the data of interest blinded.

Over time Atlas contributors have developed simulations of many parts of the standard model. Hence, constructing the model required for the background involves selecting and combining relevant simulations. If there is no simulation for some aspect that you require, then it must be requested or you may have to build it yourself. Similarly, if there is no relevant data of interest in the experimental data repository, it must be requested from subsequent capture from the detectors when LHC is next fired up in the appropriate energy levels. This comes from a completely separate team running the (non-virtual) experiment.

The development of the background is approximately a one person-year activity as it involves the experimental design, the design and refinement of the model (software simulations), the selection of methods and tuning to achieve the correct signature (i.e., get the right data), verify the model (observe expected outcomes when tested), and dealing with errors (statistical and systematic) that arise from the hardware or process. The result of the Background phase is a model approved by the collaborative to represent the background required by the experiment with the signal region blinded. The model is an "application" that runs on the Atlas "platform" using Atlas resources - libraries, software, simulations, and data much drawing on the ROOT framework, CERN's core modeling and analysis infrastructure. It is verified by being executed under various testing conditions.

This is an incremental or iterative process each step of which is reviewed. The resulting design document for the Top Quark experiment was approximately 200 pages of design choices, parameter settings, and results - both positive and negative! All experimental data and analytical results are probabilistic. All results have error bars; in particle physics they must be at least 5 sigma to be accepted. This explains the year of iteration in which analytical models are adjusted, analytical methods are selected and tuned, and results reviewed by the collaboration.

The next step is the actual virtual experiment. This too takes months. You might be surprised to find that once the data is un-blinded (i.e., synthetic data is replaced in the region of interest with experimental data), the experimenter, often a PhD candidate, gets one and only one execution of the "verified" model over the experimental data.

Hopefully this portion of a use case illustrates that DIA is a complex but critical tool in scientific discovery used with a well-defined understanding of veracity. It must stand up to scrutiny that evaluates if the experiment - consisting of all models, methods, and data with probabilistic results and error bounds better than 5 sigma – is adequate to be accepted by Science or Nature as demonstrating that the hypothesized correlation is causal.

4. Research For An Emerging Discipline

The next step in this research to better understand the theory and practice of the emerging discipline of Data

Science; to understand and address its opportunities and challenges; and to guide its development, is given in its definition. Modern Data Science builds on conventional Data Science and on all of its constituent disciplines required to design, verify, and operate end-to-end DIA activities, including both data management and DIA processes, in a DIA ecosystem for a shared community of users. Each discipline must be considered with respect to which it contributes to investigating phenomena, acquiring new knowledge, and correcting and integrating new with previous knowledge. Each operation must be understood with respect to which correctness, completeness, and efficiency can be estimated.

This research involves identifying relevant principles and techniques. Principles concern the theories that are established formally, e.g., mathematically, and possibly demonstrated empirically. Techniques involve the application of wisdom [21], i.e., domain knowledge, art, experience, methodologies, practice, often called best practices. The principles and techniques, especially those established for conventional Data Science, must be verified and if required extended, augmented, or replaced for the new context of the Fourth Paradigm, especially its volumes, velocities, and variety. For example, new departments at MIT, Stanford, and the University of California, Berkeley, are conducting such research under what some are calling *21st Century Statistics*.

A final, stimulating challenge is what is called *meta-modelling* or *meta-theory*. DIA, and more generally Data Science, is inherently multi-disciplinary [10]. This area emerged in the physical sciences in the 1980s and subsequently in statistics and machine learning and is now being applied in other areas to address combining results of multiple disciplines. Analogously, meta-modelling arises when using multiple analytical models and multiple analytical methods to analyze different perspectives or characteristics of the same phenomena. This extremely natural and useful methodology, called *ensemble modelling*, is required in many physical sciences, statistics, and AI, and should be explored as a fundament modelling methodology.

Acknowledgement

I gratefully acknowledge the brilliant insights and improvements proposed by Prof Jennie Duggan, Northwestern University and Prof Thilo Stadelmann, Zurich University of Applied Sciences.

References

[1] G. Aad et al. 2012. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* 716, 1 (2012), 1–29.

[2] Accelerating Discovery in Science and Engineering Through Petascale Simulations and Analysis (PetaApps), National Science Foundation, Posted July 28, 2008.

[3] J. Bohannon, “Fears of an AI pioneer,” *Science*, vol. 349, no. 6245, pp. 252–252, Jul. 2015.

[4] G.E.P. Box. Science and Statistics. Journal of the American Statistical Association 71, 356 (April 2012), 791–799 reprint of original from 1962

[5] N. Diakopoulos. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. Tow Center. February 2014.

[6] Duggan, Jennie and Michael Brodie, *Hephaestus: Data Reuse for Accelerating Scientific Discovery*, In CIDR 2015

[7] S.J. Gershman, E.J. Horvitz, and J.B. Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science 349, 6245 (2015), 273–278.

[8] Jim Gray on eScience: a transformed scientific method, in A.J.G. Hey, S. Tansley, and K.M. Tolle (Eds.): The fourth paradigm: data-intensive scientific discovery. *Proc. IEEE* 99, 8 (2009), 1334–1337.

[9] E. Horvitz and D. Mulligan. 2015. Data, privacy, and the greater good. Science 349, 6245 (July 2015), 253–255.

[10] M.I. Jordan and T.M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 6245 (July 2015), 255–260.

[11] V. Khachatryan et al. 2012. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B* 716, 1 (2012), 30–61.

[12] Kuhn, Thomas S. The Structure of Scientific Revolutions. 3rd ed. Chicago, IL: University of Chicago Press, 1996.

[13] Mayer-Schönberger, V., & Cukier, K. (2013-03-05). Big Data: *A Revolution That Will Transform How We Live, Work, and Thi*nk. Houghton Mifflin Harcourt

[14] Miller, G. A. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information". Psychological Review 63 (2): 81–97.

[15] NIH Precision Medicine Initiative, http://www.nih.gov/precisionmedicine/

[16] D.C. Parkes and M.P. Wellman. 2015. Economic reasoning and artificial intelligence. Science 349, 6245 (July 2015), 267–272.

[17] F. Provost and T. Fawcett. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data* 1, 1 (March 2013), 51–59.

[18] Scott Spangler, et. al. 2014. 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.

[19] J. Stajic, R. Stone, G. Chin, and B. Wible. 2015. Rise of the Machines. Science 349, 6245 (July 2015), 248–249.

[20] J. W. Tukey, "The Future of Data Analysis," Ann. Math. Statist. pp. 1–67, 1962.

[21] Bin Yu, Data Wisdom for Data Science, ODBMS.org, April 13, 2015.