On Developing Data Science

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

Abstract

Understanding phenomena based on the facts – on the data – is a touchstone of data science. The power of evidence-based, inductive reasoning distinguishes data science from science. Hence, this chapter argues that, in its initial stages, data science applications and the data science discipline itself be developed inductively and deductively in a virtuous cycle.

The virtues of the *20th Century Virtuous Cycle* (aka *virtuous hardware-software cycle,* Intel-Microsoft virtuous cycle) that built the personal computer industry *(National Research Council, 2012)* were being grounded in reality and being self-perpetuating – more powerful hardware enabled more powerful software that required more powerful hardware, enabling yet more powerful software, and so forth. Being grounded in reality – solving genuine problems at scale – was critical to its success, as it will be for data science. While it lasted, it was self-perpetuating, due to a constant flow of innovation, and to benefitting all participants – producers, consumers, the industry, the economy, and society. It is a wonderful success story for 20th Century *applied science*. Given the success of virtuous cycles in developing modern technology, virtuous cycles grounded in reality should be used to develop data science, driven by the wisdom of the 16th Century proverb, *Necessity is the mother of invention.*

This chapter explores this hypothesis using the example of the evolution of database management systems over the last 40 years. For the application of data science to be successful and virtuous, it should be grounded in a cycle that encompasses industry (i.e., real problems), research, development, and delivery. This chapter proposes applying the principles and lessons of the virtuous cycle to the development of data science applications; to the development of the data science discipline itself, e.g., a data science method; and to the development of data science education; all focusing on the critical role of collaboration in data science research and management, thereby addressing the development challenges faced by the more than 150 Data Science Research Institutes (DSRIs) worldwide. A companion chapter (Brodie, 2018a), addresses essential questions that DSRIs should answer in preparation for the developments proposed here: *What is data science*? and *What is world-class data science research*?

1 Introduction

Data Science is inherently *data-* or *evidence-based analysis;* hence, it is currently an applied science. Data science emerged at the end of the 20th Century as a new paradigm of discovery in science and engineering that used *ad hoc* analytical methods to find correlations in data at scale. While there was science in each analysis, there was little science underlying data science *per se*. Data science is in its infancy and will take a decade to mature as a discipline with underlying scientific principles, methods, and infrastructure (Brodie, 2018a). This chapter describes a method by which data scientists and DSRIs might develop data science *as a science* (e.g., fundamentals - principles, models, and methods) and *as a discipline* or *an applied science* (e.g., practices in the development of data science products, as described throughout this book (Braschler et al., 2018). The method is based on the 21st Century Virtuous Cycle – a cycle of collaboration among industry, research, development, and delivery, e.g., to develop and use data science products.

The cycle and its virtues evolved from medieval roots to surface in industry including in the research and development of large-scale computer systems and applications, extended to include product development as a *research and development (R&D) cycle*; now extended to deployment in a *research, development, and delivery (RD&D) cycle*. The cycle is used extensively in academic and industrial computer science research and development, by most technology startups, and is integral to the open source ecosystem. It is used extensively in applied science and education, and increasingly in medical and scientific research and practice. We look at the lessons learned in the development of large-scale computer systems, specifically relational database systems based on a recent analysis (Brodie, 2018b),

tracing how the virtuous cycle was extended to a larger virtuous cycle of demand, research, product development, deployment, practice, and back again.

Section 2 introduces the 20th Century Virtuous R&D Cycle made famous by Microsoft and Intel. Section 3 extends the cycle to the 21st Century Virtuous RD&D Cycle, illustrated using the mutual development of database management system (DBMS) research and products; and extends the cycle to education. Section 4 builds upon this blueprint and applies it to three aspects of data science: concrete data products, the discipline itself, and data science education; and concludes by looking forward. Section 5 illustrates previous themes with lessons learned in the development of data science and DSRIs, and exposing commonly reported data science facts as pure myths. Section 6 speculates on the impacts of data science, both benefits and threats; given the projected significance data data science, there may be more profound impacts. Section 7 concludes optimistically with challenges that lie ahead.

2 20th Century virtuous cycles

The 20th Century Virtuous Cycle accelerated the growth of the personal computer industry with more powerful hardware (speed, capacity, miniaturization) that enabled more powerful software (functions, features, ease of use) that in turn required more powerful hardware (Figure 1). Hardware vendors produced faster, cheaper, more powerful hardware (i.e., chips, memory) fueled by Moore's Law. This led software vendors to increase the features and functions of existing and new applications, in turn requiring more speed and memory. Increasing hardware and software power made personal computers more useful and applicable to more users, thus increasing demand and growing the market that in turn, through economies of scale, lowered costs in ever-shortening cycles. But what made the cycle virtuous?

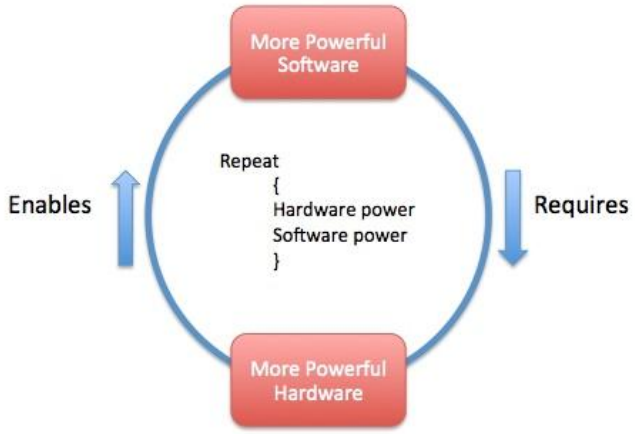


Figure 1: The Hardware-Software Cycle

The hardware-software cycle had two main virtues worth emulating. First, the cycle became self-perpetuating driven by a continuous stream of innovation - good hardware ideas, e.g., next generation chips, and good software ideas, e.g., next great applications (Figure 2). It ended in 2010 (National Research Council, 2012) when dramatic hardware gains were exhausted, the market approached saturation, and its fuel – good ideas – was redirected to other technologies. Second, all participants benefited: hardware and software vendors, customers, and more generally the economy and society through the growth of the personal computer industry and the use of personal computers. The 20th Century Virtuous Cycle was simply *hardware innovation and software innovation in a cycle*.

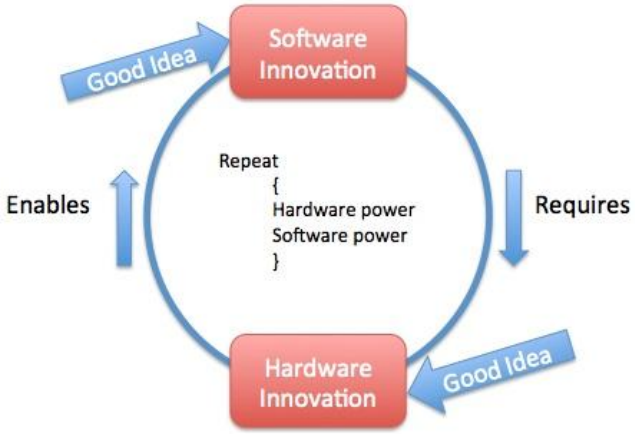


Figure 2: 20th Century Virtuous Hardware-Software Cycle

The virtuous hardware-software cycle produced hardware and software each of which developed its own R&D cycle (Figure 3). Hardware vendors and universities used the hardware (R&D) cycle to address hardware opportunities and challenges by conducting fundamental research into next generation hardware. As long as there was hardware innovation – good ideas – the hardware R&D cycle was virtuous. Similarly, software vendors used the software R&D cycle to address software challenges and opportunities in their ever-shortening cycles. This also worked well for next generation applications. However, fundamental research into next generation systems, specifically database management systems, was conducted by vendors (e.g., IBM, Software AG, Honeywell Information Systems) not by universities.

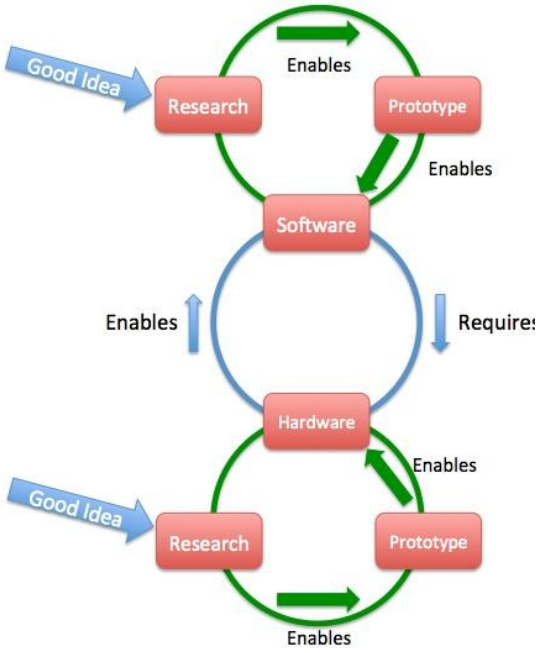


Figure 3: Software and Hardware R&D Cycles

Addressing fundamental DBMS challenges and opportunities in a university requires access to industrial-scale systems, industrial applications, and use cases (i.e., data). Until the early 1970s, universities lacked industrial experience, case studies, and resources such as large-scale systems and programming teams. At that time, Michael Stonebraker at University of California, Berkeley, began to address this gap[1]. Stonebraker and Eugene Wong built Ingres (Stonebraker et. al., 1976), a prototype industrial scale relational DBMS (RDBMS) for industrial scale geographic applications. They made the Ingres code line available as one of the first open source systems. The Ingres code line then enabled universities to conduct fundamental systems research. Ingres was the first example in a university of extending the 20th Century Virtuous Cycle to systems engineering, specifically to a DBMSs. The cycle was subsequently extended to large systems research in universities and industry. Due to the importance of the system developed in the process, it became known as the *20th Century Virtuous R&D Cycle* which simply stated is *research innovation and engineering innovation, in a cycle*.

3 21st Century virtuous research, development, and delivery cycles

3.1 The virtuous DBMS RD&D cycle

Using Ingres for industry scale geographic applications was a proof of concept of the feasibility of the relational model and RDBMSs. But were they of any value? How real were these solutions? Were relational systems applicable in other domains? These questions would be answered if there were a market for Ingres, i.e., a demand. Stonebraker, Wong, and Larry Rowe formed Relational Technology, Inc., later named the Ingres Corporation, to develop and market Ingres. Many companies have used the open source Ingres and Postgres (Stonebraker et al., 1991) code lines to produce commercial RDBMSs (Naumann, 2018) that together with IBM's DB2, Oracle, and Microsoft SQL Server now form a $55bn per year market, thus demonstrating the value and impact of RDBMSs as a "good idea" (Stonebraker, 2018a, 2018b). This extended the 20th Century Virtuous R&D Cycle to DBMSs in which DBMS research innovation led to DBMS engineering innovation that led to DBMS product innovation. DBMS vendors and universities repeated the cycle resulting in expanding DBMS capabilities, power, and applicability that in turn contributed to building the DBMS market. Just as the hardware-software cycle became virtuous, so did the DBMS R&D cycle. First, research innovation – successive good ideas – led to engineering innovation that led to product innovation. This cycle continues to this day with the emergence of novel DBMS ideas especially with the new demands of Big Data. Second, all participants benefit: vendors, researchers, DBMS users, and more generally the economy using data management products and the growth of the data management industry. Big Data and data science follow directly in this line.

A wonderful example of *necessity being the mother of invention* is the use of abstract data types as the primary means of extending the type system of a DBMS and providing an interface between the type systems of a DBMS and its application systems; arguably Stonebraker's most significant technical contribution. To build an RDBMS based on Ted Codd's famous paper (Codd, 1970), Stonebraker and Wong obtained funding for a DBMS to support Geographic Information Systems. They soon discovered that it required point, line, and polygon data types and operations that were not part of Codd's model. Driven by this necessity, Stonebraker chose the emerging idea of abstract data types to extend the built-in type system of a DBMS. This successful innovation has been a core feature of DBMSs ever since. Abstract data types is only one of many innovations that fed the 40-year-old virtuous necessity-innovation-development-product cycle around Ingres and Postgres.

In all such cycles, there is a natural feedback loop. Problems (e.g., recovery and failover), challenges, and opportunities that arose with relational DBMS products fed back to the vendors to improve and enhance the products while more fundamental challenges (e.g., lack of points, lines, and polygons) and opportunities went back to university and vendor research groups for the next cycle of innovation.

[1] Stonebraker's DBMS developments coincided with the emergence of the open source movement. Together they created a virtuous cycle that benefited many constituencies - research, DBMS technology, products, applications, users, and the open source movement resulting in a multi-billion-dollar industry. Hence, this example warrants a detailed review as lessons for the development of data science.

Indeed, modern cycles use frequent iteration between research, engineering, and products to test or validate ideas, such as the release of beta versions to find "bugs".

Stonebraker, together with legions of open source contributors, extended the 20th Century Virtuous R&D Cycle in several important dimensions to become the *21st Century Virtuous Research, Development, and Delivery Cycle*. First, in creating a commercial product he provided a compelling method of demonstrating the value and impact of was claimed as a "good idea" in terms of demand in a commercial market. This added the now critical delivery step to become the research-development-delivery (RD&D) cycle. Second, as an early proponent of open source software on commodity Unix platforms he created a means by which DBMS researchers and entrepreneurs have access to industrial scale systems for RD&D. Open source software is now a primary method for industry, universities, and entrepreneurs to research, develop, and deliver DBMSs and other systems. Third, by using industry scale applications as use cases for proofs of concept, he provided a method by which research prototypes could be developed and demonstrated to address industrial scale applications. Now benchmarks are used for important industrial scale problems as a means of evaluating and comparing systems in industrial scale contexts. Fourth, and due to the above, his method provided means by which software researchers could engage in fundamental systems research, a means not previously available that is now a critical requirement for large-scale systems research.

The RD&D cycle is used to develop good research ideas into software products with a proven demand. Sometimes the good idea is a pure technical innovation, e.g., a column store DBMS*: queries will be much faster if we read only the relevant columns*! that led to the Vertica DBMS (Stonebraker et al., 2005). More often it is a "pain in the ass" (PIA) problem, namely a genuine problem in a real industrial context for which someone will pay for the development of a solution. Paying for a solution demonstrates the need for a solution and helps fund its development. Here is a real example: A major information service company creates services, e.g., news reports, by discovering, curating, de-duplicating, and integrating hundreds of news wire reports from data items that are dirty, heterogeneous, and highly redundant, e.g., over 500 reports of a US school shooting in 500 different formats. Due to the Internet, as the number of news data sources soared from hundreds to hundreds of thousands, the largely manual methods would not scale. This PIA problem led to Tamr[2], a product for curating data at scale.

The RD&D cycle is the process underlying applied science. The RD&D cycle – an applied science method – becomes virtuous as long as there is a continuous flow of good ideas and PIA problems that perpetuate it (Figure 4).

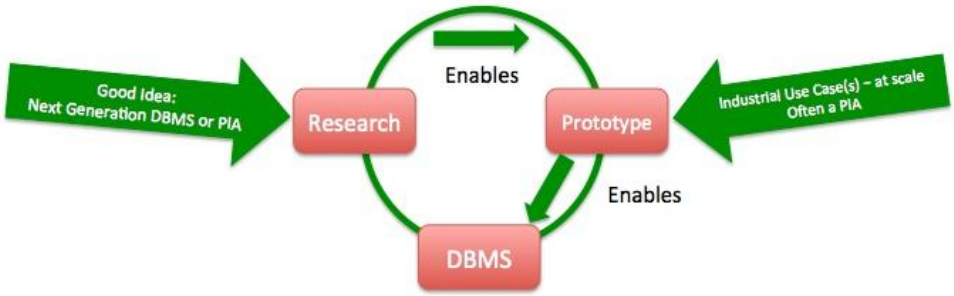


Figure 4: Virtuous DBMS RD&D Cycle

Stonebraker received the 2014 A. M. Turing Award - "the Nobel prize in computing" – "For fundamental contributions to the concepts and practices underlying modern database systems" (ACM, 2015)[3]. Concepts mean good research ideas – DBMS innovations. Practice means taking DBMS

[2] Tamr.com provides tools and services to discover and prepare data at scale, e.g., 100,000 data sources, for Big Data projects and data science.

[3] The RDBMS RD&D cycle was chosen to illustrate the theme of this chapter, as it is one of the major achievements in computing.

innovations across the virtuous RD&D cycle to realize value and create impact. Following the cycle produced the open source Ingres DBMS that resulted in the Ingres DBMS product, and the Ingres Corporation with a strong market, i.e., users who valued the product. Stonebraker refined and applied his method in eight subsequent academic projects and their commercial counterparts: Ingres (Ingres) (Stonebraker et al., 1976), Postgres (Illustra) (Stonebraker et al., 1991), Mariposa (Cohera), Aurora (StreamBase), C-Store (Vertica) (Stonebraker et al., 2005), Morpheus (Goby), H-Store (VoltDB), SciDB (Paradigm4), and Data Tamer (Tamr), with BigDAWG Polystore and Data Civilizer currently in development. The concepts and practice of this RD&D cycle are a formula for applied science of which Stonebraker's systems are superb examples[4](Stonebraker, 2018a):

Repeat {
Find somebody who is in pain
Figure out how to solve their problem
Build a prototype
Commercialize it
}

The systems research community adopted open source methods and extended the cycle to all types of systems resulting in a *21st Century Virtuous RD&D Cycle* for systems that transformed academic systems research to deliver greater value for and higher impact in research, industry, and practice.

The *21st Century Virtuous Research, Development, and Delivery Cycle* is simply *research innovation, engineering innovation, and product innovation, in a cycle*. As we will now see, its application and impacts go well beyond systems RD&D.

3.2 The critical role of research-industry collaboration in technology innovation

Virtuous RD&D cycles require researchers-industry collaboration that mutually benefits research *and* industry. Industry often needs insight into challenges for which they may not have the research resources. More commonly, industry faces PIA problems for which there are no commercial solutions. As discussed in section 4.2 this is precisely the case for data science today. Most US enterprises have launched data science efforts most of which fail as few in industry understand data science nor can hire data scientists. But lets return to understanding the cycles before applying them to data science.

It is common that industry may not be aware of PIA problems that lurk below the surface. For example, all operational DBMSs, more than 5m in the USA alone, decay due to their continuous evolution to meet changing business requirements. While database decay is a widely-known pattern, it has not been accepted as a PIA problem since there is little insight into its causes, let alone technical or commercial solutions. Recent research (Stonebraker et al., 2016a, 2016b, 2017) proposes both causes and solutions that will be realized only with industrial scale systems and use cases with which to develop, evaluate, and demonstrate that the proposed "good ideas" actually work! Insights into causes and solutions came exclusively through a research-industry collaboration between MIT and B2W Digital, a large Brazilian retailer.

Industry gains in RD&D cycles in several ways. First, industry gains insight into good ideas or challenges being researched. Second, industry gets access to research prototypes to investigate the problem in their environment. Third, if successful, the prototype may become open source[5] available to industry to apply and develop, potentially becoming a commercial product. Fourth, industry can gain ongoing benefits from collaborating with research such as facilitating technology transfer and indicating to customers, management, and investors its pursuit of advanced technology to improve its products and services. Finally, a PIA industry problem may be resolved or a hypothesized opportunity may be realized.

---

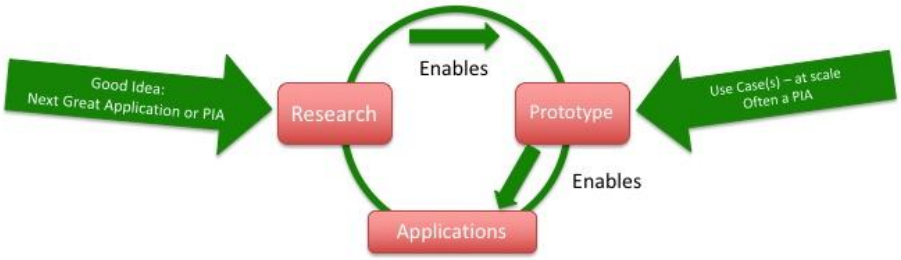
[4] Don't let the pragmatism of these examples hide the scientific merit. Computer science was significantly advanced by fundamental principles introduced in each of the systems mentioned.

[5] Open source is not required for research-industry collaborations; however, open source can significantly enhance development, e.g., Apache Spark's 42m contributions from 1,567 contributors; and impact, e.g., used by over 1m organizations, due in part to free downloads.

Industry collaboration is even more critical for research, especially for research involving industrial-scale use cases. Researchers need access to genuine, industrial scale opportunities or, more often challenges, that require research that is beyond the capability or means of industry to address, and to real use cases with which to develop, evaluate, and demonstrate prototype solutions. Scale is important as "the devil is in the details" that arise in industrial-scale challenges and seldom in toy use cases. Through collaboration, research can understand and verify the existence and extent of a problem or the likelihood and potential impact of a good idea by analyzing them in a genuine industrial context. Is the problem real? Is a solution feasible? What might be the impact of the solution? This is precisely what is needed in data science for both researchers and industry.

Ideally, collaboration occurs in a continuous RD&D cycle in which research and industry interact to identify and understand problems, opportunities, and solutions. It is virtuous if all participants benefit and as long as problems and opportunities arise. Such research–industry collaborations are better for technology transfer than conventional marketing and sales (Stonebraker, 2018a,b).

By the mid-2000s startups worldwide used a version of the 21st Century Virtuous RD&D Cycle (Figure 5) as their development method as a natural extension of the open source ecosystem. An obvious example is the World Wide Web, that spawned an enormous number of apparently odd innovations. Who knew that a weird application idea like Twitter, a 140-character message service, would become a thing (weaponized by a US president)? Or Snapchat, an image service where images self-destruct? The virtuous RD&D cycle was used on a much grander scale in the World Wide Web and in Steve Jobs' iPhone both of which went from self-perpetuating to viral and in so doing changed our world. These projects were



developed, and continue to be developed, with extensive industry collaboration driven by good – sometimes weird – ideas, novel applications, and PIA problems to be proven at scale. One might argue that the 21st Century Virtuous RD&D Cycle is one of the most effective development methods.

Figure 5: 21st Century Virtuous RD&D Cycle

3.3 The role of innovation in RD&D Cycles

The virtues of the RD&D cycle apply to data science. First, data science should be grounded in reality by using industrial-scale challenges, opportunities, and use cases to drive the cycle to develop and validate solutions and products to prove value and impact. Second, it should be made self-perpetuating by ensuring a constant flow of innovation, especially in its emerging state – good ideas, challenges, PIA problems, and opportunities – with the result that the methods and results improve, thus benefiting all participants: producers, consumers, the industry, the economy, and society. Innovative ideas perpetuate the cycle, the best innovations accelerate the cycle.

As illustrated in Figure 6, innovation is required in each stage, for the cycle to be virtuous – to self-perpetuate. There is a two-way flow between cycle stages. Technology, e.g., a data science platform, transfers down (→) the cycle in the form of research results, prototypes, and products, while requirements transfer up (←) the cycle in the form of use cases, PIA problems, opportunities, challenges, and user requirements. Innovation – good ideas – can enter anywhere in the cycle, but must continuously enter for the cycle to self-perpetuate.

The cycle also applies to education - understanding *How* each stage works and educating participants in its successful application. For data science education, understanding *How* stages work leads to data science theories in research, to data science architectures and mechanisms in engineering, to data science products in development, and to data science applications in practice. Education also benefits from a two-way flow between theories in research, architectures in engineering, products in development, and use cases in practice. Innovation – good ideas – can enter anywhere in the cycle.

Education in an established domain such as DBMSs involves understanding the principles and techniques and *How* they work. Innovation for education across the cycle concerns innovation not only in data science *per se* but also in education – how data science is taught and understood. Research and technology transfer across the cycle requires innovation in each stage. The cycle is more dynamic and powerful in an emerging domain such as data science. Each stage in data science is in its infancy; hence each stage in research could involve developing, generalizing, and integrating the current results in that stage – principles, platforms, products, and practice. Applying virtuous cycle principles to data science means grounding the work in a real challenge, e.g., drug discovery in cancer research (Spangler et al., 2014), with industrial-scale challenges and opportunities to drive the cycle, real use cases to develop and validate solutions, and products to determine value and impact. In the cancer case just cited, innovation occurred, i.e., Spangler et. al. developed a domain-specific data science method that was subsequently generalized to be more domain independent (Nagarajan et al., 2015), and the mechanisms used to further verify the results are now more widely applied in data science.

Activity	**Research**		**Engineering**		**Development**		**Delivery**
Result	**Publication**		**Prototype**		**Product**		**Application / Use Case**
Applied to Technology							
20th C. hardware-software R&D cycle	Innovation	←→			Innovation		
20th C. Infrastructure / Systems RD&D cycle	Innovation	←→	Innovation	←→	Innovation		
21st C. RD&D cycle	Innovation	←→	Innovation	←→	Innovation	←→	Innovation
Applied to Research, and Education, and Technology Transfer							
Education	How	←→	How	←→	How	←→	How
Research & Technology Transfer	Innovation	←→	Innovation	←→	Innovation	←→	Innovation

Figure 6: The Flow of Good Ideas in Virtuous Cycles

3.4 Establishing Causality: A Critical Challenge

Due to the critical problems to which data science is being applied, e.g., IBM Watson is in the business of recommending medical treatments, it is critical that accurate likelihoods of outcomes be established. One of the greatest challenges of data science is doing just that – establishing accurate

estimates of probabilistic outcomes and error bounds for those results, to which we now turn our attention.

The objective of the 21st Century Virtuous RD&D Cycle is to continuously produce technology and applications that are grounded in reality, namely that produce products that create value, or even a market of such products that have positive practical, economic, and social impacts. For example, there is a market for data science-based systems that automate aspects of online retailers supply chain, e.g., automatically buying hundreds of thousands of products to meet future sales while not overstocking. In 2015 the cost of overstocking was approximately $470bn and of understocking $630bn worldwide (Economist, 2018d). Normal economics and the marketplace are the mechanisms for demonstrating value and measuring impact. Determining value and impact is far from simple. Most technology such as DBMSs and products such as Microsoft Office have immense value and impact with continuously growing, multi-billion dollar markets. Data science-based products have the potential for great contributions to individuals, organizations, society, and the economy. Like most technology, data science holds equal potential for positive and negative impacts. Disliking a Netflix data-science-driven movie recommendation may waste half an hour of your time. Unfortunately, substantial negative consequences are being discovered in data science applications, such as ethical problems in parole sentencing used extensively in the USA (O'Neil, 2016). What might be the impact of data-driven personalized medicine treatment recommendations currently being pursued by governments around the world?

Consider that question given that *Why Most Published Research Findings Are False* (Ioannidis, 2005) has been the most referenced paper in medical research since 2005. Data science currently lacks robust methods of determining likelihood of and error bounds for predicted outcomes, let alone how to move from such correlations to causality. While mathematical and statistical research may be used to address probabilistic causality and error bounds, consider the research required to address ethical and societal issues such a sentencing.

The scientific principles that underlie most research also underlie data science. Empirical studies report causal results while data science cannot. Data science can accelerate the discovery of correlations (Brodie, 2018a). A significant challenge is to assign likelihoods and error bounds to these correlations. While the current mechanisms of the 21st Century Virtuous RD&D Cycle to measure value and impact of products worked well for simple technology products, they may not work as well for technology that is increasingly applied to every human endeavor, thus directly influencing our lives. This is a significant issue for the development and operation of data science in many domains. This is yet another class of issues that illustrate the immaturity of data science and the need for multi-disciplinary collaboration. The complex issue of causal reasoning in data science is addressed in greater detail in the companion chapter (Brodie 2018a).

4 Applying 21st Century virtuous RD&D cycles to data science

A primary benefit of the 21st Century Virtuous RD&D Cycle is to connect research, engineering, and products in a research-development-delivery cycle with the objective of being virtuous through a continuous flow of innovative, good ideas and challenging problems. The cycle has many applications. It is used extensively in computer science research in academia and industry, in startups that are building our digital world, and increasingly in medicine and science. It has been and is being used to transform education. I propose that it be used to guide and develop data science research, practice, and education.

4.1 A data science RD&D cycle example

In the mid-2000s legions of software startups applied the 21st Century Virtuous RD&D Cycle to customer facing applications. As an example, Stonebraker applied the RD&D cycle to Goby – an application that searches the web for leisure activities to provide users, e.g., tourists, with a list of distinct local, leisure activities. The “good idea” was to find all activities on the web that might be of interest to tourists. The PIA problem is that there are thousands of leisure activities with many listings that are highly redundant (i.e., replicas), dirty, often inaccurate and contradictory, and in heterogeneous formats. As is typically the case in data science analyses, more than 80% of the resources were required to discover,

deduplicate, and prepare the data, leaving less than 20% for analysis, in this case determining relevant activities. This real, industrial-scale use case led to research, Morpheus (Dohzen et al., 2006), that developed machine driven, user guided solutions to discover, clean, curate, deduplicate, integrate (a better term is unify), and present data from potentially hundreds of thousands of data sources. The "good idea" led to a PIA[6] problem that resulted in a prototype that led to a product with a commercial market that demonstrated its value and impact. Meanwhile, unanticipated challenges cycled back to Goby for product improvements and enhancements while more fundamental, research challenges went back to Morpheus. The good idea – find events on the web – was generalized from events to the data discovery and preparation of any type of information leading to further innovation that led to a new research project – Data Tamer – that in turn led to a new product – Tamr.com – and a burgeoning market in data discovery and preparation for data science (Forrester, 2017b) (Gartner, 2017c). Tamr and similar products are part of the budding infrastructures for data science, called data science platforms (Gartner, 2017a, 2017b).

The 21st Century Virtuous RD&D Cycle is being used to design, develop, and deliver data science tools and platforms. Data discovery and preparation, and data science platforms are concrete examples of this cycle in practice. Over 30 data preparation tools and 60 data science platforms are emerging (Gartner, 2017a, 2017b, 2018a, 2018b). This cycle is virtuous as long as there are continuous innovation and broad benefits. Currently, aspects of most human endeavors are being automated by means of digital tools developed to study, manage, and automate those endeavors. Data preparation tools are being developed by being applied to an increasing number of new domains, each presenting new challenges. The continuous flow of practical applications, use cases, PIA problems and other challenges contribute to the cycle being virtuous. The cycle becomes virtuous when all participants benefit. Data science tools and platforms are beginning to flip the ratio of the data-preparation to analysis resources from 80:20 to 20:80, so that data scientists can devote the majority of their time to analysis and not to plumbing. Data science practiced in a virtuous cycle is applied science at its best – producing broad value and contributing to accelerating data science practice and the development of data science *per se*.

4.2 Developing data science in practice and as a discipline

Data science is an emerging phenomenon worldwide that will take a decade to mature as a robust discipline (Brodie, 2015, 2018a). Its growth and diversity can be seen in the number (over 150) and nature of DSRIs, most of which were established after 2015. The emerging state of data science can be seen in the fact that each DSRI provides different answers to key data science questions that all DSRIs should answer (Brodie, 2018a): *What is data science*? *What is the practice of data science*? *What is world class data science research*?

The 21st Century Virtuous RD&D Cycle can guide the development and practice of data science. First, the domain is just emerging characterized by a constant flow of new ideas entering the cycle. Data science is being attempted in every human endeavor for which there is adequate data (Brodie, 2018a). Second, due to its immaturity (Brodie, 2015) data science must be grounded in reality, i.e., real data in real use cases at the appropriate scale. The cycle can be used to guide the development and work of individual data scientists and, at a greater scale, of DSRIs. Major features of the cycle are already present in most DSRIs, specifically research-industry collaboration in their research and education. Most have industry partners and collaborations for education, RD&D, for case studies, and for technology transfer. In most cases, significant funding has come from industry partners. The charter of the Center of Excellence at Goergen Institute for Data Science[7] includes collaborating with industry "to apply data science methods and tools to solve some of the world's greatest challenges in sectors including: Medicine and Health,

---

[6] Good ideas hopefully arise in answer to a PIA challenge. In this example, the good idea, finding events on the web, led to a PIA problem that was resolved with the now conventional machine driven (ML) and human guided method. The trick is a combination of good ideas and PIA challenges, leading to valuable results.

[7] http://www.sas.rochester.edu/dsc/

Imaging and Optics, Energy and the Environment, Food and Agriculture, Defense and National Security, and Economics and Finance." The mission statement of the recently launched Harvard Data Science Initiative[8] states "Applications are by no means limited to academia. Data scientists are currently key contributors in seemingly every enterprise. They grow our economy, make our cities smarter, improve healthcare, and promote civic engagement. All these activities – and more – are catalyzed by the partnership between new methodologies in research and the expertise and vision to develop real-world applications."

Applying the 21st Century Virtuous RD&D Cycle to DSRIs must recognize three factors that distinguish data science from conventional academic research that often lacks research-industry engagement. First, while core or theoretical research is equally important in both cases, DSRI resources must be allocated to applied research, technology transfer, and supporting research-industry collaboration[9]. Unlike a computer science research institute and in support of this objective, a DSRI might have a *Chief Scientific Officer* to establish DSRI-wide data science objectives, such as contributing more than the sum of its parts, and coordinating research across the many organizational units into the components of data science, e.g., principles, models, and analytical methods; pipelines and infrastructure; and a data science method, to support data science in all domains. Second, special skills, often not present in research staff, are required for research-industry engagement, the research-development-delivery cycle, and technology transfer. For example, emerging data science platforms are increasingly important for developing and conducting data science. A data science platform includes workflow engines, extensive libraries of models and analytical methods, platforms for data curation and management, large-scale computation, and visualization; that is, a technology infrastructure to support end-to-end data science workflows or pipelines. Hence, research into the development of data science platforms should be a DSRI research objective. Again, unlike a computer science research institute, a DSRI might also establish a *Chief Technology Officer* responsible for those functions including the development and maintenance of a shared data science technology infrastructure.

The third distinguishing factor is the relative immaturity of data science versus most academic disciplines; excitement and hype cloud the real state of data science. A common claim is that data science is successful, ready for technology transfer and application in most human endeavors. While there are successful data science technologies and domain-specific results, in general this impression, often espoused by vendors and enthusiasts[10], is false. While there are major successes and expert data scientists, data science is an immature, emerging domain that will take a decade to mature (Brodie, 2015, 2018a). Analysts report that most early (2010-2012) data science projects in US enterprises failed (Forrester, 2015a, 2015b) (Demirkan & Dal, 2014) (Veeramachaneni, 2016) (Ramanathan, 2016). In late 2016, Gartner reported that while most (73%) enterprises declare data science as a core objective, only 15% have deployed Big Data projects in their organization (Gartner, 2016a) with well-known failures (Lohr & Singer, 2016). This reflects confusion concerning data science and that technology analysts are not reliable judges of scientific progress.

Slow progress makes perfect sense as data science is far more complex than vendors and enthusiasts report. For example, data science platforms provide libraries of sophisticated algorithms (visualization (Matplotlib, Matlab, Mathematica); data manipulation, aggregation, and visualization (Pandas); linear algebra, optimization, integration, and statistics (SciPy); image processing and machine learning (SciKit-Learn); Deep Learning (Keras, TensorFlow, Theano); Natural Language Processing (NLTK)), that business users have significant difficulty fitting to business problems (Forrester, 2015b). There is a significant learning curve – few people understand deep learning, let alone statistics at scale – and substantial differences with conventional data analytics. *What do you mean these aren't just spreadsheets*?

---

[8] http://datascience.harvard.edu/

[9] In its emerging state, data science lacks a scientific or theoretical base. Establishing data science as a science should be a fundamental objective of data science researchers and DSRIs (Brodie, 2018a).

[10] Michael Dell, Dell CEO, predicted at the 2015 Dublin Web Summit that big data analytics is the next trillion-dollar market. IDC predicts 23.1% compound annual growth rate, reaching $48.6 billion in 2019. Forrester Research declared that "all companies are in the data business now." Gartner predicts "More than 40 percent of data science tasks will be automated by 2020" (Gartner, 2016b).

Over the next decade, research will establish data science principles, methods, practices, and infrastructure, and will address these key questions. This research should be grounded in practical problems, opportunities, and use cases. DSRIs should use the 21st Century Virtuous RD&D Cycle to direct and conduct research, practice, education, and technology transfer. Initially, they might use the R&D cycle to explore good ideas. Research-industry collaborations should be used to identify and evaluate novel data science ideas. When collaborations can identify plausible use cases or PIA problems, the research-development-delivery cycle should be used. That is, to identify research domains and directions, DSRIs should identify industrial partners with whom to collaborate to establish virtuous cycles that equally benefit researchers and industry partners. As with applied university research funding, a significant portion of data science research funding should come from industry to increase industry-research engagement and quickly identify valuable research with impact potential.

4.3 Developing data science education

Data science is one of the fastest growing subjects in education due to the demand for data scientists. Data science courses, programs, degrees, and certificates are offered by most universities and professional training institutes and are part of the mission of most DSRIs. Given the decade to maturity of data science, how should data science education programs be developed?

Just as the 21st Century Virtuous RD&D Cycle is used to transform the research, development, delivery, and use of computer systems and applications, it can also being used to transform education. The intention of the recently launched *21st Century Applied PhD Program in Computer Science*[11] at Texas State University, is for PhD level research ideas, innovations, and challenges to be developed in prototype solutions and refined and tested in industrial scale problems of industrial partners. The cycle is to be driven by industrial partners that investigate or face challenges collaboratively with the university. PhD candidates work equally in research and in industry to identify and research challenges and opportunities that are grounded in real industrial contexts; and to develop prototype solutions that are refined using industrial use cases. This educational cycle requires technology transfer from research to advanced prototypes to industry with opportunities and problems transferring, in the opposite direction, from practice to advanced development and to research. It becomes virtuous with a constant stream of "good ideas" – challenges and opportunities – and of PhD candidates in one direction, and industry PIA problems, challenges, and opportunities in the other. The primary benefits of this program are that research, teaching, and products are grounded in reality.

These ideas are not new. The Fachhochschule system (universities of applied sciences) applied virtuous cycle principles in Germany since the late 1960s, and in Austria and Switzerland since the 1990s as a graduate extension of the vocational training and apprenticeship (Berufslehre und Ausbildung) programs that have roots in mentorships and apprenticeships from the middle ages.

While the quality and intent of the European and US educational systems are the same, the systems differ. Academic universities focus on theory and applied universities focus on the application of science and engineering. Fachhochschulen usually do not grant PhDs. In addition, research in applied universities is funded differently from research in academic universities. Usually, over 80% of applied research funding comes from third parties to ensure research-industry engagement[12] and as a test of the PIA principle. Unsuccessful research is quickly identified and terminated. Dedicated government agencies provide partial funding and promote innovation and technology transfer through collaboration between industry and the applied universities. Enrollments in Fachhochschulen are soaring, indicating the demand for education grounded in reality – closely mirroring successful startup behavior. Due to the significance of, demand for, and perceived value of data science, education programs should be revisited considering adding more applied aspects to conventional research and education for data science. A good example of

---

[11] https://cs.txstate.edu/academics/phd/

[12] A similar principle applied by the funding agency in the section 5.1 story was initially considered a death knell by the DSRI and by me. It took a year for me to see the value.

this vision is the *21st Century Applied PhD Program in Data Science* at Texas State University, based on a collaborative research-industry-development-delivery model.

5 Lessons Learned

5.1 Data science and DSRI stages of development

In 2013, I was invited to join the Scientific Advisory Committee (SAC) of Ireland's Insight Center for Data Analytics, at the time one of the first and largest DSRIs, composed of four partner institutes. Since then I have actively participated on the SAC as well as on Insight's Governance Committee. Over the following years, I observed the development of Insight as a DSRI as well as the establishment of over 150 DSRIs at major institutions worldwide. Insight's development as a DSRI was not without challenges. In 2017, Science Foundation Ireland (SFI) reviewed Insight for a potential five-year funding renewal. Insight needed to tell SFI what data science was, what world class data science research was, and to measure its progress accordingly. This led me to the observation, stated to the review board, that Insight's development as a DSRI reflected the development of data science as a discipline. The most thoughtful contributors to data science fully understood that while the potential benefits for Ireland and the world were enormous, data science as a discipline was in its infancy and faced considerable scientific and organizational developmental challenges. Further, that Insight in operating for five years and in aspiring to world class data science contributions as a world class DSRI, had faced and overcome significant challenges that I had witnessed first-hand at Insight and indirectly in eight other DSRIs.

Over five years, Insight had gone through the four stages of development that younger DSRIs are just encountering. Insight is currently at stage five - a critical stage. Successful progress through the stages revolved around three fundamental issues:

- Just as the science and the scientific method are far more than experiments in a single domain, so too is data science more than data science activities in a single domain.
- Changing centuries of research behaviour to enable collaboration across disciplines in data science pipelines, as well as across academic and organizational boundaries.
- Producing, for Ireland and for data science, more than the sum of the parts, i.e., the results of individual member institutes.

The five stages are simple.

1. ***Act of creation***: An organizational decision was made to form a DSRI from independent, one might say competing, institutes with a new focus, the emerging discipline of data science. The institutes – researchers and administrators alike – in a behavioral and legal tradition of individual progress and reward – were not happy campers. Awkwardness arose.
2. ***Initial participation***: Participants continued business as usual, but expressed a willingness to participate and cooperate followed by little actual collaboration and some ingrained competitiveness. The DSRI administration soldiered on towards understanding the bigger picture that had not been defined by anyone – funders, researchers, or advisors.
3. ***Data science objectives understood – conceptually***: After a few years of successful execution of individual research efforts and attempts to understand data science, modest progress was made, especially once it was clear that funding would depend on the DSRI being more than the sum of the parts and would be measured on world-class data science, interpreted then as contributing to data science, *per se*. But what is that exactly?
4. ***Data science objectives understood – emotionally:*** Goals provide focus. Five years of funding of the now seven institutes depended on the DSRI being "more than the sum of the parts". This was not an abstract concept but required providing benefits such as accelerating discovery in specific parts of the Irish economy, educating data scientists, and economic growth in Ireland, involving not just researchers but major industrial partners. Individual researchers rose to the challenge to

propose a collaborative DSRI. By the time of the review, they had become a band of data science brothers and sisters, together with industrial partners.

5. ***Stand and deliver:*** While the DSRI will continue to produce specific data science results that are world class in specific domains, e.g., physiology, it is defining and planning contributions to data science, including data science principles, models, methods, and infrastructure (Brodie 2018a).

Many DSRIs around the world have been created, like Insight, by a higher-level organization, typically a university, to coordinate the myriad data science activities in that organization. The critical factor missing in many DSRIs, at least as viewed through their web sites, is an imperative to understand and contribute to data science *per se*, to contribute more than the sum of the contributions of the partner organizations. SFI's funding of Insight depends on contributing to data science *per se*, worded as "contributing more than the sum of the parts." This imperative is not present in many DSRI's.

5.2 Myths of applying data science in business

As often happens with new technology trends, their significance, impact, value, and adoption are exaggerated by the analysts and promoters as well as by optimists and the doomsayers. Technology analysts see their roles as reporting on new technology trends, e.g., Gartner's Hype Cycles. If a technology trend is seen as significant, investment analysts join the prediction party. Technology and investment analysts are frequently wrong as they are now with data science. Many technology trends reported by Gartner die before reaching adoption, e.g., 1980's service-oriented architectures. Some trends that are predicted as dying become widely adopted, e.g., the .com boom was reported as a failure, largely due to the .com stockmarket bubble, but the technology has been adopted globally and has led to transforming many industries. Data science is one of the most visible technology trends of the 21st Century with data scientists called "the sexiest job of the 21st Century" (Davenport, 2012) and "engineers of the future" (van der Aalst, 2014). To illustrate the extent to which data science is blown out of proportion to reality, let's consider several data science myths. A reasonable person might ask, given the scale, scope, and nature of the change of data science as a new discovery paradigm, how could anyone predict with any accuracy how valuable it will be and how it will be adopted, especially when few people, including some "experts", currently understand it (that, by the way, was myth #1).

Everyone is successfully applying data science: As reported above most (80%) early (2010-2012) data science projects in most US enterprises failed. By early 2017, while 73% of enterprises declare data science as a core objective, only 15% have deployed it. In reality, AI/data science is a hot area, with considerable, perceived benefit. Hence many companies are exploring it. However, such projects are not easy and require ramping up of rare skills, methods, technologies. It is also difficult know when and how to apply the technology and to appropriately interpret the results. Hence, most initial projects are highly unlikely to succeed but are critical to gain the expertise. Applying AI/data science in business will have major successes (10%) and moderate successes (40%) (Gartner, 2016a). Most companies are and should explore AI/data science but be prepared for a significant learning curve. Not pursuing AI/data science will likely be an advantage to your competitors.
Reality: organizations perceiving advantages should explore data science and prepare for a learning curve.

Data science applications are massive: While scale is a definitive characteristic of Big Data and data science, successful applications can be small and inexpensive. The pothole example (Brodie 2018a) was a very successful launch of now flourishing startups in the emerging domain of autonomous vehicles. It was based on building and placing small, inexpensive (~$100) motion detectors in seven taxis. It started with the question shared by many business, *What is this data science stuff*? It was a pure exploration of data science and not to find a solution to a PIA problem. As data science matures, we see that the critical characteristics of a data analysis are determined by the domain and the analytical methods applied. Volume is one characteristics that must meet statistical requirements but even GB or TB may be adequate and can be handled readily by laptops.

Reality: data science can be applied on modest data sets to solve interesting, small problems.

Data science is expensive: Famous, successful data analytics (Higgs Boson, Baylor-Watson cancer study, LIGO, Google, Amazon, Facebook) often require budgets at scale (e.g., massive processing centers, 100,000 cores, 1,000s of analysts); however, data analytics even over relatively large data volumes can be run on desktops using inexpensive or free open source tools and the cloud. Businesses can and should conduct initial explorations like the pothole analyses at negligible cost.
Reality: small players with small budgets can profit from data science.

Data science predicts what will happen: Otto, a German retailer orders 200,000 SKUs fully automated. Above we cited predictions of trillions of dollars in related savings worldwide. However, the results of good data analytics are at best probabilistic with error bounds. This is somewhat similar to science (scientific method) but is typically less precise with lower probabilities and greater error bounds due to inability of applying the controls that are applied in science. Businesses should explore the predictive power of data science but with the full understanding of its probabilistic and error prone nature. Otto and the supply chain industry constantly monitors and verifies results and adjusts as needed or, like H&M, you might end up with a $4.3bn overstock (New York Times, 2018).
Reality: predictions are probabilistic and come with error bounds.

Data science is running machine learning over data: Machine learning is highly visible in popular press accounts of data science. In reality, one must select from thousands of AI and non-AI methods and algorithms depending on the phenomenon being analyzed and the characteristics of the data. What's more, as reported above, while algorithm selection is critical, 80% of the resources including time for a data analysis is required just to find and prepare the data for analysis.
Reality: as there is no free lunch (Wolpert, 1997), there is no single methodology, algorithm or tool to master to do successful data science, just as it is in science.

AI/data science is disrupting and transforming conventional industries and our lives: This widely reported myth (Marr, 2017) (Chipman, 2016) makes eye catching press but is false. There is ample evidence that AI/data science is being applied in every human endeavor for which adequate data is available such as reported throughout this book. The list of impacted industries is long: mechanical engineering & production of industrial goods (shop floor planning, robotics, predictive maintenance); medicine (personalized health); commerce/trade (e-commerce, online business, recommenders); hospitality (demand planning and marketing via analytics, pricing based on customer analytics); transportation (ride-sharing/hailing); automotive (self-driving cars, e-mobility); services (new business models based on data); and many more. In reality, five industries have been massively disrupted by digital innovation —music, video-rental, publishing (books, newspapers), taxicabs, and retailing (predominantly clothing). They are in the process of being transformed, e.g., the Spotify business model is an example of transformation in music; Uber's is in taxicabs, but the process takes years or decades. However, the vast majority of industries are currently unaffected. If an industry is being transformed, it is reflected in the stock market, e.g., a price-earnings ratio of less than 12 is generally forecast imminent collapse. According to that rule of thumb, Ford and GM's price-earnings ratio of 7 suggest disruption and transformation if not collapse possibly due to electric vehicles (EVs) such as Tesla and ride-sharing/hailing. There are no such indications for the other "conventional industries" (Economist, 2017).
Reality: almost all conventional industries are impacted, but only few are disrupted.

It's all about AI: Current popular and even scientific press suggests that AI is one of the hottest and potentially most significant technologies of the 21st century. AI is sometimes referred to as an object as in "an AI is used to …". Without doubt AI and specifically machine learning (ML) and deep learning (DL) have been applied to a wide range of problems with significant success and impact as described above. It is very probable that ML will be applied much more extensively with even greater success and impact. However, like most "hot" technical trends, the press characterization is a wild exaggeration – a myth. First

of all, AI is a very broad field of research and technology that pursues all forms of intelligence exhibited by a machine (Russell & Norvig, 2010). ML is one of perhaps 1,000 AI technologies. Second, until we understand ML, its application will be limited. The current, very successful ML technologies arose in the early 2000s from a previously unsuccessful technology, neural networks. Amazingly ML, augmented by massive data sets and high-performance computing, has been applied to images, sentences, and data to appear to identify entities that are meaningful to humans, e.g., pizzas, cats, trains on tracks, and cluster those meaningful entities based on similarities meaningful to humans. We have no idea why there is a correlation between the results of an ML analysis and meaning understood by humans. Considerable research is being invested in understanding such reasoning, but it is far from mature. As a result, the use of ML in the European Community is restricted by the GDPR law. Finally, the successful application of ML is proportional to the data to which it is applied, typically ML works most effectively on massive data sets. Massive data analysis requires high performance computing, one of the critical components that moved neural networks from failure to success. Hence, most naïve misuses of the term AI should be replaced with the specific AI technology, e.g., ML, plus data plus high-performance computing. This sounds remarkably like data science.

Reality: AI is a key component, amongst many others, that are necessary to conduct data science; AI does not perform miracles; in many cases "AI" should be replaced by the technologies used to support analytical workflows.

6 Potential impacts of data science

The development of data science involves not just the science, technology, and applications, it also involves the opportunities, challenges, and risks posed by the applications of data science. Hence, I now briefly review some potential benefits and threats of applying data science, many of which have been reported in the popular press. However, popular press descriptions of hot technical topics and their impacts are usually to be taken with a grain of salt, especially concerning AI and data science that are not well understood by some experts.

In the early 2010s Big Data was the hot technology topic. In 2018, AI and its power was the hot topic, not unreasonably as Sundar Pichai, Google CEO, said that AI will have a "more profound" impact than electricity or fire (Economist, March 2018a). Consider it a matter of terminology. Big Data, on its own is of no value. Similarly, without data AI is useless. The hot technical topics that surface in the media are equally attributable to AI, massive data, and powerful computation. In what follows, as above, I refer to applications of that combination as AI/data science. Yet even those three terms are not adequate to achieve the hot results since data science depends also on domain knowledge and more, but this will suffice for the following discussion.

According to Jeff Dean, director of Google's AI research unit, Google Brain, more than 10m organizations "have a problem that would be amenable to a machine-learning solution. They have the data but don't have the experts on staff." (Economist, March 2018b) That is, the potential impacts of AI/data science will have broad applicability.

As with a new, powerful technology, society, e.g., legislation, is seldom able to keep up with its impacts. In the case of AI/data science, let's consider its impacts on our daily lives, both the benefits and the threats, of a multi-billion-dollar industry that currently has almost no regulations to restrain its application.

6.1 Benefits

Google's youthful founders, Sergey Brin and Larry page, defined its original vision "to provide access to the world's information in one click" and mission statement "to organize the world's information and make it universally accessible and useful" with the famous motto "Don't be evil". Indeed, they *almost* succeeded beyond their wildest dreams. The entire world benefits from instant access to much of the world's information. We went from this utopian view of the potential of AI/data science to

one in which Google, in 2015, dropped its famous motto from its code of conduct. I address some shortcomings, such as use and protection of personal information, in the next section.

It is infeasible to list here the many benefits of AI/data science, so let's consider two examples, medicine and autonomous vehicles. A data-science based medical analysis can compare a patient's mammogram with 1m similar mammograms in seconds to find potential causes and treatments that were most effective for the conditions present in the subject mammogram. Similar analyses and achievements are being made with our genetic code to identify the onset of a disease and effective treatment plans based on millions of similar cases, something no human doctor could possibly do on their own.

Autonomous vehicles depend on AI/data science. It is commonly projected that autonomous vehicles will radically reduce the 1m annual traffic deaths per year worldwide, pollution and traffic congestion while shortening travel times, freeing us up for a better quality of life. The impacts could be far greater than those of the automobile. But how will autonomous vehicles change the world? One factor to consider is that currently the average car sits parked 95% of the time. What might be the impacts of autonomous vehicles on real estate, roads, automobile manufacturing, and employment?

Most benefits of technology harbor unanticipated threats. For example, autonomous vehicle results can be applied in many domains, e.g., autonomous weapons are used by 80 countries including the USA that has over 10,000[13]. Let's consider a few threats posed by AI/data science.

6.2 Threats

On May 6, 2010 the US stock market crashed. In the 2010 Flash Crash, over a trillion dollars in value was lost and the indexes (Dow Jones Industrial average, S&P 500, Nasdaq Composite) collapsed (Dow Jones down ~1,000 points, 9% in value). Within 36 minutes the indexes and value largely but not entirely rebounded. This was due in part to algorithmic trading that operates 60% of trading in US exchanges, and in part to the actions of Navinder Singh Sarao, a trader who the U.S. Department of Justice convicted for fraud and market manipulation[14]. Algorithmic trading is a data science-based method of making trades based on complex economic and market analysis based on potentially all trades ever transacted. This was not a threat. It was a reality and a harbinger of similar threats.

How might AI/data science threaten employment? Consider the potential impact of autonomous vehicles on America's 4m professional drivers (as of 2016, US Bureau of Labor Statistics). Robots will impact a vastly larger number of jobs. McKinsey Global Institute estimated that by 2030 up to 375m people could have their jobs "automated away" (Economist 2018c). These examples are the tip of the AI/data science unemployment iceberg. The Economist (Economist 2018f) and the Organisation for Economic Co-operation and Development (OECD) (Nedelkoska, et. al. 2018), estimate that over 50% of all jobs are vulnerable to automation.

An insidious threat is bias in decision making. Our lives are increasingly determined by algorithms. Increasing machine learning and other sophisticated algorithms are used to make decisions in our lives, in our companies, in our careers, in our education, and in our economy. These algorithms are developed with models that represent the significant features of the problem being addressed. No one but the developers see the code, fewer people actually understand the code. So, what is in the code? Are race, sex, or a history of past behaviour significant and *acceptable* features in parole sentencing? Are these algorithms biased against certain types of individuals? ProPublica proved that parole sentencing is indeed biased against blacks (Angwin et al., 2016). The twelve vendors of the systems that the US government uses for sentencing refuse to release their code for inspection. Ironically, ProPublica proved their case using data science. They collected and analyzed sentencing data to prove with high confidence that the systems were inherently biased. This has led to the algorithmic accountability movement in the legal community.

---

[13] Do you trust this computer? http://doyoutrustthiscomputer.org

[14] https://www.justice.gov/opa/pr/futures-trader-pleads-guilty-illegally-manipulating-futures-market-connection-2010-flash

In many countries, tech companies, e.g., Apple, Alphabet (Google parent), Microsoft, Amazon, Facebook, Alibaba, and Tencent, know more about us and can predict our behaviour better than we can. In some countries, the government takes this role (e.g., China's social credit system). Over the past decade, there has been increasing concern for personal information. Legislation to govern the use and privacy of personal information (General Data Protection Regulation[15] (GDPR)) was enacted in Europe only in May 25, 2018. US congressional hearings only began in early 2018 prompted by the alleged illegal acquisition of 87m Facebook profiles by Cambridge Analytica (CA), described below.

The power and growth of the seven companies mentioned above, the largest companies in the world by market capitalization, is directly attributable to AI/data science. Their average age is less than ten years in contrast to average age of 141 years of the legacy companies that they are supplanting from the top ten largest companies. These tech leaders vastly outspend the largest legacy companies in research and development, e.g., Apple's 2017 $22.6bn R&D investment was twice that of non-tech Johnson & Johnson, established in 1886. It is frequently argued (Lee, 2017) (Economist, 2018e) that the power of AI/data science is such that the country that dominates the field will wield disproportionate economic and ultimately political power worldwide, i.e., will monopolize not just AI/data science but areas of the economy for which it is a critical success factor. Currently China and the USA are the leaders by far. However, the playing field is beginning to favor China. The power and development of AI solutions is heavily dependent on vast amounts of data. Increasing restrictions on data, such as privacy legislation mentioned above, will significantly inhibit US AI companies while there is little or no such limitations by the Chinese government that itself collects massive data on its citizens.

It may seem dramatic, but data science has allegedly been used to threaten democracy. Alexander Nix, the now-suspended CEO of now-insolvent CA, claimed to have applied a data science-based methodology, psychometrics, to influence political opinion. Nix reported that it was used to influence the outcomes of political elections around the world including the 2016 British EU referendum, aka Brexit referendum, in favor of leaving, and the 2016 US election in favor of Donald Trump. Psychometrics is based on a physiological profiling model from Cambridge and Stanford Universities. For the US election, CA illegally and legally acquired up to 5,000 data points each of 230m Americans to develop a detailed profile of every American voter. Profiles were used to send "persuasion" messages (e.g., on gun rights) targeted to and nuanced for the weaknesses and preferences of individual voters. CA activities were first reported in 2015 and resurfaced in January 2017 when Trump took office. It wasn't until April 2018 that CA's actions in the 2016 US election were considered for prosecution. Notwithstanding CA's illegal actions and potentially violating American democratic principles, CA's data-science method appears to have been very effective and broadly applicable, e.g., being applied in targeted, 1-on-1 marketing. Such methods are allegedly being used by governments, e.g., in the Chinese social credit system and in Russian interference with the 2016 US election. This genie is out of the bottle.

6.3 More profound questions

A more profound question is: Will these advanced technologies enhance or replace man? In *Homo Deus* (Harari, 2016), the author Yuval Noah Harari, hypothesizes that the human race augmented by advanced technologies, specifically AI/data science, will transform homo sapiens into a new species. Just as homo sapiens surpassed and replaced Neanderthals, so will humans augmented with machines surpass homo sapiens without automation. Could you compete or survive without automation? This is well beyond considering the impacts of data science. Or is it? In 2018 there were multiple attacks on the very foundations of democracy (see above). At the TED 2018 conference, Jaron Lanier, virtual reality creator, suggested that, using data, social networks had become behaviour modification networks. Harari speculated that just as corporations use data now, so too could dictatorships use data to control populations.

Technological progress is never solely positive, e.g., automation that eliminates waste due to optimized supply chains. Progress is relative to our expectations, e.g., computers will eliminate most

[15] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

human drivers thereby reducing road accidents by 95%. In this case, the cost of saving lives is a loss of jobs. The greatest impacts of technology are seldom foreseen, e.g., the redistribution of populations from cities to suburbs due to the mobility offered by automobiles. What might be the impact of a machine beating humans playing Jeopardy?

The Future of Life Institute[16] was established "To catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges." Its motto is: "Technology is giving life the potential to flourish like never before... or to self-destruct. Let's make a difference."

7 Conclusions

Data Science is potentially one of the most significant new disciplines of the 21st Century, yet it is just emerging, poses substantial challenges, and will take a decade to mature. The potential benefits and risks warrant developing data science as a discipline and as a method for accelerated discovery in any domain for which adequate data is available. That development should be grounded in reality following the proverb: *Necessity is the mother of invention*. This chapter proposes a long standing, proven development model.

Innovation in computing technology has flourished through three successive versions of the virtuous cycle. The 20th Century Virtuous Cycle was hardware innovation and software innovation in a cycle. The 20th Century Virtuous R&D Cycle was research innovation and engineering innovation in a cycle. The emerging 21st Century Virtuous RD&D Cycle is research innovation, engineering innovation, and product innovation in a cycle. While innovation perpetuates the cycle, it is not the goal. Innovation is constantly and falsely heralded as *the* objective of modern research. Of far greater value are the solutions. Craig Vintner – a leading innovator in genetics – said, "Good ideas are a dime a dozen. What makes the difference is the execution of the idea." The ultimate goal is successful, efficient solutions that fully address PIA problems or major challenges, or that realize significant, beneficial opportunities. Data science does not provide such results. Data science accelerates the discovery of probabilistic results within certain error bounds. It usually does not produce definitive results. Having rapidly reduced a vast search space, to a smaller number of likely results, non-data science methods, typically conventional methods in the domain of interest are used to produce the definitive results. Once definitive results are achieved, the data science analysis can be converted to a product, e.g., a report, inventory replenishment, etc. however, the results of such a product must be monitored as conditions and data can change constantly. For more on this see (Meierhofer et. al., 2018).

The principles and objectives of the 21st Century Virtuous RD&D Cycle are being applied in many domains beyond computer science, startups, education, and data science. In medicine it is called translational medicine (STM, 2018) in which healthcare innovation and challenges go across the *benchside/research-bedside-community*[17] cycle, delivering medical innovations to patients and communities more rapidly than conventional medical practice and taking experience and issues back for research and refinement. The US National Institutes of Health (NIH) established The National Center for Advancing Translational Sciences in 2012 for this purpose and is increasingly requiring its practice in NIH funded research programs. In the broader scientific community, such activities are called translational science and translational research, e.g., (AJTR, 2018) (Fang & Casadevall, 2010). The RD&D cycle is now incorporated in all natural science and engineering research funded in Canada[18].

Data science researchers and DSRI leaders might consider the 21st Century Virtuous RD&D Cycle to develop and contribute to data science theory, practice, and education.

---

16 https://futureoflife.org

17 The US National Institutes of Health support of translational medicine in which the research process includes testing research (benchside) results in practice (bedside) to speed conventional clinical trial methods.

18 Dr. Mario Pinto, President of the Natural Sciences and Engineering Research Council of Canada, in 2017 announced that a research-development-delivery method was to be used in all NSERC funded projects.

Acknowledgement
Thanks to Dr. Thilo Stadelmann, Zurich University of Applied Sciences, Institute for Applied Information Technology in the Swiss Fachhochschule system, for insights into these ideas; and to Dr. He H. (Anne) Ngu, Texas State University, for insights into applying these principles and pragmatics to the development of Texas State University's 21st Century Applied PhD Program in Computer Science.

8 References

ACM (2015). Michael Stonebraker, 2014 Turing Award Citation[19], Association of Computing Machinery, April 2015

AJTR (2018). American Journal of Translational Research, e-Century Publishing Corporation[20].

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks, ProPublica, May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Brodie, M.L. (2015). 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.

Brodie, M.L. (2018a). What is Data Science? to appear in (Braschler, et. al. 2018).

Brodie, M.L. (Ed.) (2018b). Making Databases Work: The Practical Wisdom of Michael Stonebraker, A.M. Turing Book Series, ACM Books, Forthcoming Summer 2018.

Chipman, I., (2016). How data analytics is going to transform all industries, Stanford Engineering Magazine, February 13, 2016.

Codd. E.F. (1970). A relational model of data for large shared data banks. Commun. ACM 13, 6 (June 1970), 377-387.

Davenport, T. H., and Patil, D.J., (2012). "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review Vol.* 90, no. 10 (October 2012).

Demirkan, H. & Dal, B. (2014). The Data Economy: Why do so many analytics projects fail? Analytics Magazine, July/August 2014.

Dohzen, T., Pamuk, M., Seong, S. W., Hammer, J., & Stonebraker, M. (2006). Data integration through transform reuse in the Morpheus project (pp. 736–738). ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006.

Economist (March 2018a). GrAIt expectations, Special Report AI in Business, The *Economist*, March 31, 2018.

Economist (March 2018b). External Providers: Leave it to the experts, Special Report AI in Business, The *Economist*, March 31, 2018.

Economist (March 2018c). The future: Two-faced, Special Report AI in Business, The *Economist,* March 31, 2018.

Economist (March 2018d). Supply chains: In algorithms we trust, Special Report AI in Business, The *Economist,* March 31, 2018.

---

[19] http://amturing.acm.org/award_winners/stonebraker_1172121.cfm

[20] http://www.ajtr.org

Economist (March 2018e). America v China: The battle for digital supremacy: America's technological hegemony is under threat from China, The *Economist,* March 15, 2018.

Economist(2018f). A study finds nearly half of jobs are vulnerable to automation, *The Economist*, April 24, 2018.

Economist (2017). Who's afraid of disruption? The business world is obsessed with digital disruption, but it has had little impact on profits, The Economist September 30, 2017.

Fang, F. C. & Casadevall, A. (2010). Lost in Translation-Basic Science in the Era of Translational Research, Infection and Immunity, vol. 78, no. 2, pp. 563–566, Jan. 2010.

Forrester (2015a). Brief: Why Data-Driven Aspirations Fail, Forrester Research, Inc., October 7, 2015

Forrester (2015b). Predictions 2016: The Path from Data to Action for Marketers: How Marketers Will Elevate Systems of Insight. Forrester Research, November 9, 2015

Forrester (2017b). The Forrester WaveTM: Data Preparation Tools, Q1 2017, Forrester, March 13, 2017

Gartner G00326555 2018a Magic Quadrant for Analytics and Business Intelligence Platforms 26 February 2018.

Gartner G00326456 2018b Magic Quadrant for Data Science and Machine-Learning Platforms 22 February 2018

Gartner G00301536 (2017a). 2017 Magic Quadrant for Data Science Platforms, 14 February 2017.

Gartner G00326671 (2017b). Critical Capabilities for Data Science Platforms, Gartner, June 7, 2017.

Gartner G00315888 (2017c) Market Guide for Data Preparation, Gartner, 14 December 2017

Gartner G00310700 (2016a). Survey Analysis: Big Data Investments Begin Tapering in 2016, Gartner, September 19, 2016

Gartner G00316349 (2016b). Predicts 2017: Analytics Strategy and Technology, Gartner, Report G00316349, November 30, 2016.

Harari, Y.N. (2016). Homo Deus: a brief history of tomorrow, Random House, 2016

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False? PLOS Medicine, 2(8), e124.

Lee, Kai-Fu, The Real Threat of Artificial Intelligence, New York Times, June 24, 2017

Lohr, S. & Singer, N. (2016) How Data Failed Us in Calling an Election, New York Times, November 10, 2016.

Marr, B., (2018). How Big Data Is Transforming Every Business, In Every Industry, Forbes.com, November 21, 2017.

Meierhofer, J., Stadelmann, T., & Cieliebak, M. (2018). Data Products. In: Braschler, M., Stadelmann, T., & Stockinger, K. (Editors). Applied Data Science - Lessons Learned for the Data-Driven Business. Springer (to appear).

Nagarajan, M. 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.

National Research Council (2012). The New Global Ecosystem in Advanced Computing: Implications for U.S. Competitiveness and National Security. Washington, DC: The National Academies Press.

Naumann, F. (2018). Genealogy of Relational Database Management Systems[21], Hasso-Plattner Institüt, Universität, Potsdam.

Nedelkoska, L. & G. Quintini (2018), "Automation, skills use and training", *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, http://dx.doi.org/10.1787/2e2f4eea-en.

New York Times (2018). H&M, a Fashion Giant, Has a Problem: $4.3 Billion in Unsold Clothes, New York Times, March 27, 2018.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY, USA.

Olson, M. (2018). Stonebraker and open source, to appear in (Brodie 2018)

Palmer, A. (2018) How to create & run a Stonebraker Startup-- The Real Story, to appear in (Brodie 2018)

Piatetsky, G. (2016). Trump, Failure of Prediction, and Lessons for Data Scientists, KDnuggets, November 2016.

Ramanathan, A. (2016). The Data Science Delusion, Medium.com, November 18, 2016.

Spangler, S. 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.

STM (2018). Science Translational Medicine, a journal of the American Association for the Advancement of Science.

Stonebraker, M. (2018a). How to start a company in 5 (not so) easy steps", to appear in (Brodie 2018b)

Stonebraker, M. (2018b). Where Do Good Ideas Come from and How to Exploit Them? to appear in (Brodie 2018b)

Stonebraker, M., & Kemnitz, G. (1991). The Postgres Next Generation Database Management System. Communications of the ACM, 34(10), 78–92.

Stonebraker, M., Abadi, D. J., Batkin, A., Chen, X., Cherniack, M., Ferreira, M., et al. (2005). C-store: a column-oriented DBMS, In Proceedings of the 31st international conference on Very large data bases, 2005.

Stonebraker, M., Castro Fernandez, R., Deng, D., & Brodie, M.L. (2016a). Database Decay and What to do about it. Comm. ACM 60, 1 (December 2016), 10-11.

Stonebraker, M., Deng, D., & Brodie, M. L. (2016b). Database Decay and How to Avoid It (pp. 1–10). Proceedings of the IEEE International Conference on Big Data, Washington, DC.

Stonebraker, M., Deng, D., & Brodie, M. L. (2017). Application-Database Co-Evolution: A New Design and Development Paradigm. New England Database Day, (pp. 1–3) January 2017

Stonebraker, M., Wong, E., Kreps, P., & Held, G. (1976). The Design and Implementation of INGRES. ACM Transactions on Database Systems, 1(3), 189–222.

van der Aalst, W. M. P. (2014). Data Scientist: The Engineer of the Future. In K. Mertins, F. Bénaben, R. Poler, & J.-P. Bourrières (Eds.), (pp. 13–26). Presented at the Enterprise Interoperability VI, Cham: Springer International Publishing.

Veeramachaneni, K. (2016). Why You're Not Getting Value from Your Data Science, Harvard Business Review, December 7, 2016.

[21] https://hpi.de/naumann/projects/rdbms-genealogy.html

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82.