Necessity is the mother of invention: On Developing Data Science

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

Draft: October 20, 2017

1 Abstract

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. Given the success of virtuous cycles used to develop modern technology, we argue that data science applications and the discipline itself be driven by the wisdom of the 16th Century proverb, *Necessity is the mother of invention.*

The virtues of the *20th Century Virtuous Cycle* (aka *virtuous hardware-software cycle,* Intel-Microsoft virtuous cycle) that built the personal computer industry *[15]*, were being grounded in reality and 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 absolutely 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 all participants benefitting - producers, consumers, the industry, the economy, and society. It is an excellent example of the success of 20th Century *applied science*.

The cycle and its virtues evolved from medieval roots to surface 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. We look in detail at the lessons learned in development of large-scale computer systems, specifically relational database systems, tracing how the virtuous cycle was extended to a larger virtuous cycle of demand, research, product development, deployment, and practice.

A companion chapter addresses the question What is data science?[7] This chapter applies virtuous cycle principles and lessons to the development of data science as a discipline and as a method; of data science education; and of data science practice focusing on the critical role of collaboration in research and management, e.g., in Data Science Research Institutes (DSRI).

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 (see 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?

The hardware-software cycle had two main virtues. First, the cycle was not a perpetual motion machine; it 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 (see Figure 2). It ended in 2010 [15] 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, i.e., automation. The 20th Century Virtuous Cycle was simply *hardware innovation and software innovation in a cycle*.

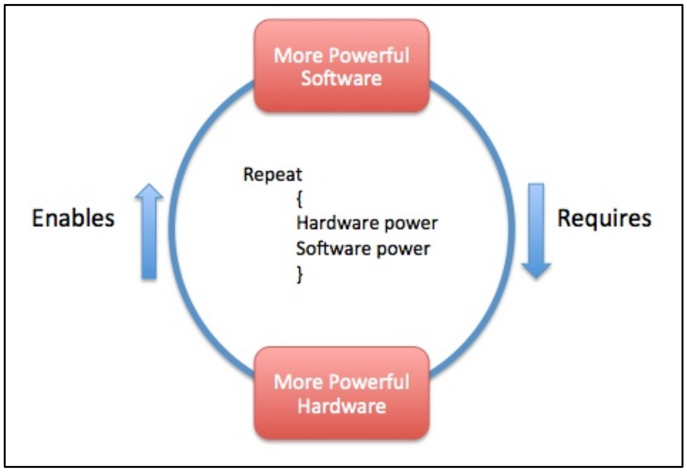


Figure 1: The Hardware-Software Cycle

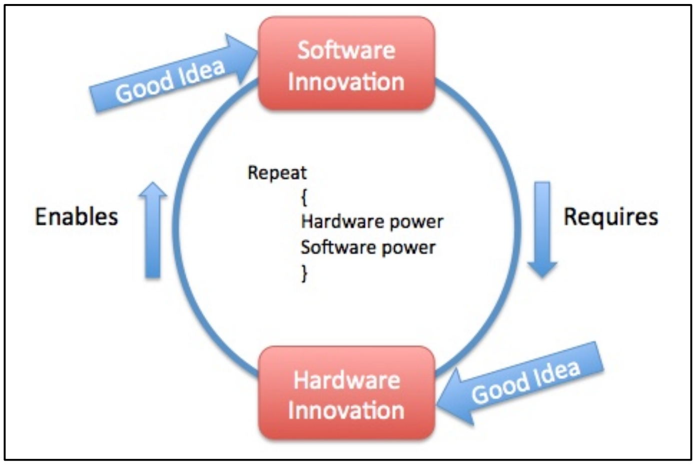


Figure 2: 20th Century Virtuous Hardware-Software Cycle

The virtuous hardware-software cycle produced hardware and software each of which had its own R&D cycle (see 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 worked well for next generation applications. However, fundamental research into next generation systems, specifically database management systems (DBMS), was conducted by vendors (e.g., IBM, Software AG, Honeywell Information Systems) not by universities.

Addressing fundamental DBMS challenges and opportunities in a university requires technical artifacts such as industrial-scale systems, and 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, Berkley, began to address this gap[1]. Stonebraker and Eugene Wong built Ingres[32], a prototype industrial scale relational DBMS (RDBMS) for industrial scale geographic applications. They made the Ingres code line

[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, DBMSs, products, applications, users, and the open source movement, and contributed to a multi-billion-dollar industry. This cycle warrants a detailed review as lessons for the development of data science.

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*.

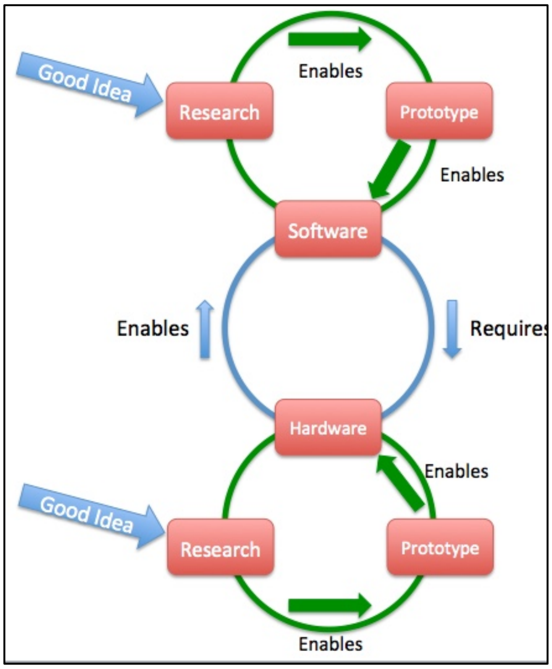


Figure 3: Software and Hardware R&D Cycles

3 The 21st Century Virtuous Research, Development, and Delivery Cycle

3.1 The Virtuous DBMS RD&D Cycle

Using Ingres for industry scale geographic applications was a proof of concept of the relational model and relational database systems. 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. 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[25] code lines to produce commercial relational DBMSs (RDBMS) [16] that together with IBM's DB2, Oracle, and Microsoft's SQL Server now form a $35B per year market, thus demonstrating the value and impact of RDBMSs as a "good idea"[30]. 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.

A wonderful example of *necessity being the mother of invention* is 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[8] , 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 virtuous necessity-innovation–development-product cycle around Ingres and Postgres. The value of these innovations can be seen in the millions of copies of open source Ingres and Postgres downloaded for DBMS research and development, and their use in commercial DBMSs, illustrated in the Genealogy of Relational Database Systems[16] and described in[18].

In all such cycles, there is a natural feedback loop. Problems, challenges, and opportunities that arose with relational DBMS products fed back to the vendors to improve and enhance the products while more fundamental challenges and opportunities went back to university and vendor research groups for the next cycle of innovation. 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 what started 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. 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[26]. 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 dirty, heterogeneous, and redundant news wire reports. 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, 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 (see Figure 4). This is the process of applied science.

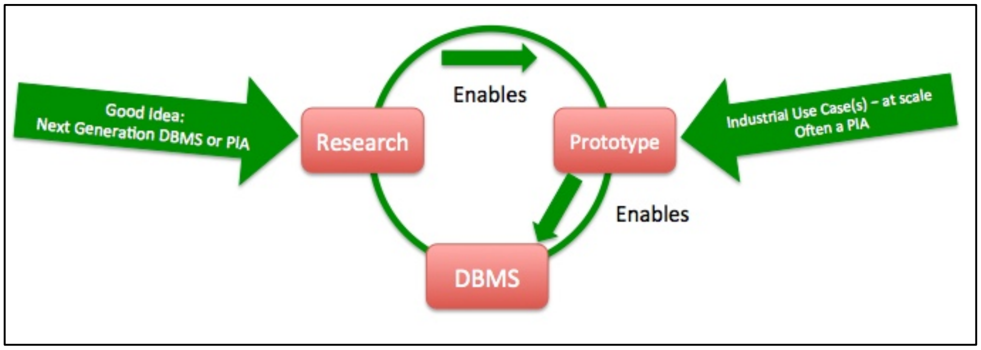

Good Idea:
Next Generation DBMS or PIA
Research
Enables
Prototype
Industrial Use Case(s) – at scale
Often a PIA
Enables
DBMS

Figure 4: Virtuous DBMS RD&D Cycle

Stonebraker received the 2014 A. M. Turing Award "For fundamental contributions to the concepts and practices underlying modern database systems."[1] Concepts mean good research ideas – DBMS innovations. Practice means taking DBMS innovations across the virtuous RD&D cycle to realize value and create impact. Following the cycle for the open source Ingres DBMS that resulted in the Ingres DBMS product, Ingres Corporation with a strong market, Stonebraker refined and applied his method in eight subsequent projects: Postgres (Illustra), Mariposa (Cohera), Aurora (StreamBase), C-Store[26] (Vertica), 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[2].

Stonebraker, ever succinct[3], characterizes his RD&D cycle as[30]:

```
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 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. It is also common that industry may not be aware of PIA problems that lurk below the surface. For example, all operational DBMSs decay due to the continuous evolution of application requirements. This is common in 5M operational databases in the USA alone. 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 [28][29][27] 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!

Industry gains 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[4] 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.

Industry collaboration is even more critical for research, especially for large-scale systems research. Researchers need access to genuine industrial scale opportunities or more often challenges that require research that is beyond the capability or means of industry, and to real use cases with which to develop, evaluate, and demonstrate prototype solutions. Through collaboration, research can understand and

[2] 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.

[3] Mike's 19 words compare with 1,057 words used above, but who's counting?

[4] 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, over 1M organizations, due in part to free downloads.

verify the existence and extent of a problem or the likelihood and potential impact of a good systems idea by analyzing them in a genuine industrial context. Is it real? Is a solution feasible? What might its impact be? Second, research requires real, industrial scale use cases in which the problem or opportunity manifests. Real use cases are critical in understanding the problem and in developing a potential solution, and to develop, test, and validate a prototype solution. Use cases provide a basis for a proof of concept of the solution.

Ideally, collaboration occurs in a continuous RD&D cycle in which research and industry interact to identify and understand problems, opportunities, and solutions throughout the cycle. 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.

By the mid-2000s most startups worldwide used a version of the 21st Century Virtuous RD&D Cycle (see Figure 5) as their development method as a natural extension of the open source ecosystem. "Good ideas" are cool[5]. But 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 the images self-destruct? The virtuous RD&D cycle was used on a much grander scale in Tim Berners-Lee's CERN WorldWideWeb project that became 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.

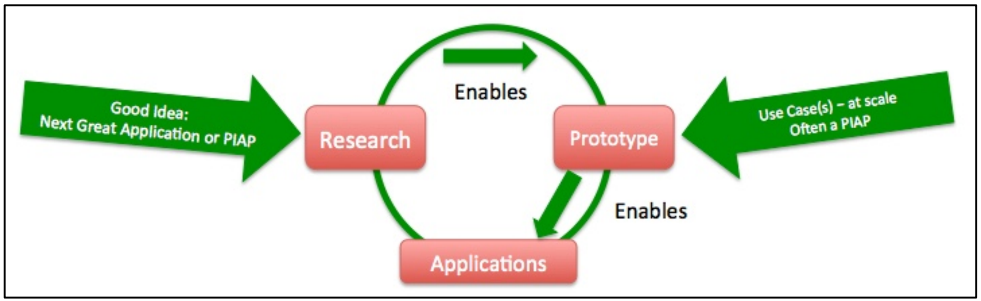


Figure 5: 21st Century Virtuous RD&D Cycle

4 Applying the 21st Century Virtuous RD&D Cycle to Data Science

4.1 The Research-Development-Delivery Cycle

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.

4.2 A Data Science 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 it 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 a tourist. The PIA problem is that there are hundreds or thousands of leisure activities with many listings that are highly redundant (i.e., replicas), very 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

[5] In demand.

and prepare the data, leaving less than 20% for the analysis, in this case determining relevant activities. This real, industrial-scale use case led to research, Morpheus[10], that developed machine driven, user guided solutions to discover, clean, curate, de-duplicate, integrate, and present data from potentially hundreds of thousands of data sources. The "good idea" led to a PIA 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 challenges went back to Morpheus. The good idea was generalized from events to data discovery and preparation or 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 bourgeoning market in data discovery and preparation for data science. Tamr and similar products are part of the budding infrastructures for data science, called data science platforms.

This cycle is virtuous as long as there are continuous innovation and broad benefits. In the 21st Century, aspects of most human endeavors are being automated by means of digital tools developed to study, manage, and automate those endeavors. For more than two decades, this has produced a continuous flow of not just good ideas but amazing ideas that have perpetuated the RD&D cycle. Hence, the 21st Century Virtuous RD&D Cycle is being used to build our digital world. Second, all participants benefit: application developers and users, the related industries, the economy, and society. This is applied science at its best – contributing to science and producing broad value.

4.3 Developing Data Science as a Discipline

Data science is an emerging worldwide phenomenon that will take a decade to mature as a robust discipline[6][7]. Its growth and diversity can be seen in the number and nature of DSRIs most of which were established within the past two years. The emerging state of data science can be seen in the fact that each DSRI provides different answers to key questions (addressed in [7]): What is data science? What is the practice of data science? What is world class data science research? What does a data scientist do? and What are the technology requirements of data science?

The 21st Century Virtuous RD&D Cycle provides a 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. Second, due to its immaturity data science must be grounded in reality, i.e., real use cases at the appropriate scale. The cycle could be used to guide the development and work of DSRIs. Major features of the cycle are present in most DSRIs, specifically research-industry collaborations in their research and education structure and operations. Most have industry partners and collaborations for education, RD&D, for case studies, and for technology transfer. The charter of the Center of Excellence at Goergen Institute for Data Science 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, 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 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[6]. Unlike a computer science research institute and in support of this objective, a DSRI might have a *Chief Scientific Officer* to coordinate research into the components of data science, e.g., principles, models, and analytical methods; data science pipelines, 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-

[6] 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[7].

industry engagement, the research-development-delivery cycle, and technology transfer. For example, data science platforms are a fundamental requirement for developing and conducting data science. A data science platform includes workflow engines, extensive libraries of 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 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 establish a *Chief Technology Officer* responsible for those functions including the development and maintenance of a shared data science technology infrastructure. Third, is the relative immaturity of data science versus most academic research; hype clouds 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[7], 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 [6] [12]. Analysts report that most early (2010-2012) data science projects in US enterprises failed [4][9][13][20][22]. In late 2016, they reported that while most (73%) enterprises declare data science as a core objective, only 15% have deployed big data projects in their organization [33] with well-known failures [14]. 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 (machine learning, deep learning, principal component analysis, statistics), that business users have significant difficulty fitting to business problems[20]. There is a significant learning curve (Few people understand deep learning, let alone statistics at scale.) and substantial differences from conventional data analytics.

Over the next decade, research will establish data science theories, methods, and practices, and address the key questions. This research should be grounded in the 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 could use the R&D cycle to explore good ideas. Research-industry collaborations should be used to identify and evaluate whether novel data science ideas are “good”. When the 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.

4.4 Developing Data Science Education

Data science is one of the fastest growing areas 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 is also being used to transform education. The intention of the recently launched *21st Century Applied PhD Program in Computer Science* 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 context; and to develop prototype solutions that are refined using

[7] 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”[21]. Simultaneously, Gartner, Forrester, and other analysts predicted in 2016 that most enterprise data science projects will fail though 2019 reflecting some confusion concerning data science. Technology analysts are seldom reliable judges of scientific progress.

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 PIA industry problems 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 and Austria in the 1950s, and in Switzerland in 1995 as a graduate extension of the vocational training and apprenticeship (Berufslehre and 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. Fachhochschules 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 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 Fachhochschules are soaring indicating the demand for education grounded in reality – closely mirroring successful startup behavior. Due to the significance of data science, these systems should be revisited including adding more applied aspects to conventional research and education for data science. Imagine a *21st Century Applied Program in Data Science* based on a collaborative research-industry research-development-delivery model.

Similar objectives are being applied in many domains beyond computer science, startups, and education to medicine where it is called translational medicine[23] in which healthcare innovation and challenges go across the *benchside/research-bedside-community* 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[8] and translational research, e.g., see[11][2].

4.5 Making the RD&D Cycle Work for Data Science

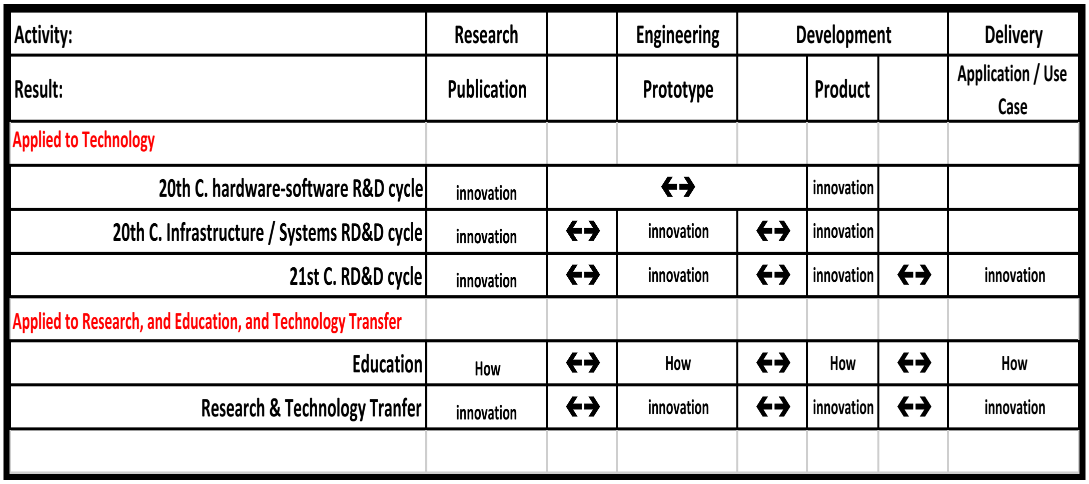
The virtues of the original virtuous cycle should apply to data science via the RD&D cycle. First, data science should be grounded in reality by using industrial-scale challenges, opportunities, and use cases to drive the cycle and to develop and validate solutions and products to prove value and impact. Second, it should be made self-perpetuating by ensuring a constant flow of innovation - good ideas, challenges – PIA problems, and opportunities - to drive the cycle so that all participants benefit - 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 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, problems, and user requirements. Innovation – good ideas – can enter anywhere in the cycle, and must enter for the cycle to perpetuate.

For education, the activity in each stage is to understand *How* the stage works. For research and technology transfer innovation is required for progress. For example, the results for data science education and research are data science theories in research, data science architectures and mechanisms in engineering, data science products in development, and data science applications in practice. These results require two-way flows between theories in research, architectures in engineering, products in development, and use cases in practice. Again, innovation – good ideas – can enter anywhere in the cycle.

[8] Dr. Mario Pinto, President of the Natural Sciences and Engineering Research Council of Canada recently endorsed the research-development-delivery method to be used in all NSERC funded projects.

Education in an established domain such as DBMSs involves understanding what exists and *How* they work. Innovation for education across the cycle concerns innovation not in data science *per se* but 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 – theories, 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[24], 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. The cycle can be used to guide data science with mechanisms to validate research value and potential impact and to filter out less promising directions. Due to the critical problems to which data science will be applied, one of its greatest challenges is establishing probabilistic causality and error bounds of results, to which we now turn our attention.



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 Tranfer	innovation	↔	innovation	↔	innovation	↔	innovation

Figure 6: Virtuous Cycles

4.6 Establishing Causality: A Critical Challenge

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 value to create a market and ultimately have positive practical, economic, and social impacts. Normal economics and the market place are the mechanisms for proving 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. Emerging products and data science, as a domain, 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, an ethical example is in crime and parole sentencing used extensively in the USA [17]. What might be the impact of data-driven personalized medicine treatment recommendations being pursued by governments around the world? Consider that question given that *Why Most Published Research Findings Are False* [12] 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 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. 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.

5 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 old proverb: *Necessity is the mother of invention*. Happily, there is a wonderful 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. Having rapidly reduces the scale of the search space, means conventional to the domain are used to produce the desired solutions.

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

Acknowledgement

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

6 References

[1] 2014 Turing Award Citation, Association of Computing Machinery, April 2015.

[2] American Journal of Translational Research, e-Century Publishing Corporation.

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

[4] Brief: Why Data-Driven Aspirations Fail, Forrester Research, Inc., October 7, 2015

[5] Brodie, M.L. (Ed.), Making Databases Work: The Works of Michael R. Stonebraker, A.M. Turing Book Series, ACM Books, Forthcoming Summer 2018.

[6] Brodie, M.L., 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.

[7] Brodie, M.L., What is Data Science? to appear in[3]

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

[9] Demirkan, H. and B. Dal, The Data Economy: Why do so many analytics projects fail? Analytics Magazine, July/August 2014

[10] Dohzen, T., Pamuk, M., Seong, S. W., Hammer, J., & Stonebraker, M. 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.

[11] Fang, F. C. and A. Casadevall, "Lost in Translation--Basic Science in the Era of Translational Research," *Infection and Immunity*, vol. 78, no. 2, pp. 563–566, Jan. 2010.

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

[13] Kalyan Veeramachaneni, Why You're Not Getting Value from Your Data Science, Harvard Business Review, December 7, 2016.

[14] Lohr, S. and N. Singer, How Data Failed Us in Calling an Election, New York Times, November 10, 2016.

[15] 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.

[16] Naumann, F., Genealogy of Relational Database Management Systems, Hasso-Plattner Institut, Universität, Potsdam.

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

[18] Olson, M., Stonebraker and open source, to appear in [5]

[19] Palmer, A., How to create & run a Stonebraker Startup-- *The Real Story*, to appear in [5]

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

[21] Predicts 2017: Analytics Strategy and Technology, Gartner, Report G00316349, November 30, 2016.

[22] Ramanathan, A., The Data Science Delusion, Medium.com, November 18, 2016.

[23] Science Translational Medicine, a journal of the American Association for the Advancement of Science.

[24] 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.

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

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

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

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

[29] 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

[30] Stonebraker, M., How to start a company in 5 (not so) easy steps", to appear in [5]

[31] Stonebraker, M., Where Do Good Ideas Come from and How to Exploit Them? to appear in [5]

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

[33] Survey Analysis: Big Data Investments Begin Tapering in 2016, Gartner, September 19, 2016

[34] The Forrester WaveTM: Data Preparation Tools, Q1 2017, Forrester, March 13, 2017

[35] Trump, Failure of Prediction, and Lessons for Data Scientists, by Gregory Piatetsky