The Evolution of Data Science: A New Mode of Knowledge Production

Article *in* International Journal of Knowledge Management · November 2018

DOI: 10.4018/IJKM.2019040106

2 authors:

Jennifer Priestley
Kennesaw State University
64 PUBLICATIONS **576** CITATIONS

SEE PROFILE

Robert J Mcgrath
University of New Hampshire
34 PUBLICATIONS **173** CITATIONS

SEE PROFILE

International Journal of Knowledge Management
Volume 15 • Issue 2 • April-June 2019

The Evolution of Data Science:
A New Mode of Knowledge Production

Jennifer Priestley, Kennesaw State University, USA
Robert J. McGrath, University of New Hampshire, USA

ABSTRACT

Is data science a new field of study or simply an extension or specialization of a discipline that already exists, such as statistics, computer science, or mathematics? This article explores the evolution of data science as a potentially new academic discipline, which has evolved as a function of new problem sets that established disciplines have been ill-prepared to address. The authors find that this newly-evolved discipline can be viewed through the lens of a new mode of knowledge production and is characterized by transdisciplinarity collaboration with the private sector and increased accountability. Lessons from this evolution can inform knowledge production in other traditional academic disciplines as well as inform established knowledge management practices grappling with the emerging challenges of Big Data.

KEYWORDS

Big Data, Computer Science, Data Science, Knowledge Management, Knowledge Production, Statistics, Transdisciplinarity

INTRODUCTION

The terms “big data”, “data science” and “analytics” have pervaded the global common speak over the past decade. While populist in many cases, these terms are rooted in the real practice of being able to measure and analyze phenomena in larger amounts, faster and with a longer and more robust historical perspective, all facilitated by technological advances and the lower cost of data storage. Data, once defined by a numerical representation of some measurement, has today evolved into an atomic unit that can be captured – that is measured, seen or heard – and thus extracted, analyzed and converted into information and ultimately into new knowledge. What began only a few years ago as a growing swell of the data ocean has become a tsunami of impacts into everyday life, or the “datafication” of the economy (Dumont, 2016).

This datafication has resulted in many organizations sprinting to better leverage the data they collect and capture the data they do not. The argument that knowledge, as a summation of data through the knowledge management pyramid (Ackoff, 1989), is the only sustainable source of competitive advantage is arguably more relevant today than when it was first posited (Drucker, 1995). It has also led many companies to declare that they are, in fact, data and information organizations more so than

DOI: 10.4018/IJKM.2019040106

they are purveyors of the products they sell (e.g. Capital One (Dee, 2016), Alibaba (Liyakasa, 2015) and Ford (Blanco, 2016)). Cities too are becoming "smarter" with data-driven innovations geared at efficient energy consumption, optimized traffic and parking, and the promotion of green and healthy practices. And individuals are becoming more data driven, with many exploring opportunities by an ever increasing "quantified self"; a concept related to the self-tracking of any number of physical, behavioral, social and many other phenomena by individuals (Swan, 2013). A revolution, or perhaps evolution, to be sure.

An unexpected consequence of these rapid (r)evolutionary changes has been the emergence of the ubiquitous and pervasive "talent gap" – the term used to describe the challenge of organizations to find people with the necessary skills to extract and analyze massive amounts of data (structured and unstructured) to generate meaningful information. Simply put, the demand for these skills has materialized so rapidly, traditional sources of supply for new talent (i.e., colleges and universities) have been ill-equipped to develop and train talent at the scale and pace demanded.

The issues related to the emergence of data science and the associated talent gap have implications for larger conversations related to organizational knowledge management. Jennex (2017) recognized the role of Big Data in the revised knowledge management pyramid. The traditional pyramid first presented by Ackoff (1989) established the framework that organizational wisdom derives from knowledge, information, and finally from data. In the revised pyramid, Jennex places a finer lens on the lowest level of the pyramid by calling out incremental layers between information and reality. These new layers include "Data", defined as "discrete facts…that can be stored in a database" (Jennex & Bartczak, 2013), "Big Data", defined as data that is "too big, too fast or too hard for existing tools to process" (Madden, 2012), and "IoT", defined as a sensor network of networks with devices continually generating vast amounts of data and facilitating the evolving definition of what data even is. This evolution in thinking from a simplistic single layer at the base of the pyramid to a more detailed treatment of data within the knowledge management pyramid increases the resolution of the lens through which reality can be detected.

It is the concepts, tools, and algorithms around "data science" that will enable a sustainable organizational approach to the translation of the layers of data into information, knowledge and ultimately to organizational wisdom/intelligence. However, where those organizational knowledge activities meet societal ones and who addresses those "fault" lines become an issue as data sources become more democratized and real time (Spender, 2007; Money & Cohen, 2018).

These types of issues have led many in academia to consider the conversations around "data science" more formally. Is this truly a new field of study, or is data science simply an extension or specialization of a discipline that already exists, such as statistics or computer science, or mathematics? The answers to these questions are not trivial and have implications for both academics as well as practitioners engaged in addressing the challenges in knowledge production and management related to the emergence of Jennex's more detailed treatment of data within the knowledge management pyramid.

A Brief History of Data Science

The term "data science" has been traced back to computer scientist Peter Naur in 1960 (Naur, 1992), but "data science" also has evolutionary seeds in statistics. In 1962, the famed statistician John W. Tukey wrote:

For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have … come to feel that my central interest is in data analysis… data analysis is intrinsically an empirical science. (Tukey, 1962)

The fields of data manipulation have grown largely through methods in mathematics, statistics and computer science during this period, with research from Peter Naur, who published "Concise Survey of Computer Methods" in 1974; Gregory Piatetsky-Shapiro who organized and chaired the

first Knowledge Discovery in Databases (KDD) workshop in 1989 and Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, who published "From Data Mining to Knowledge Discovery in Databases" in 1996. A reference to the term "data science" as a discipline within statistics was made in the proceedings of the Fifth Conference of the International Federation of Classification Societies in 1996. In 1997, during his inaugural lecture as the H. C. Carver Chair in Statistics at the University of Michigan, Jeff Wu actually called for statistics to be renamed "data science" and statisticians to be renamed "data scientists".

Since the beginning of the 21st century, data stockpiles have expanded exponentially, largely due to advents in processing and storage that is both efficient and economical at scale, leading to the drive to collect, analyze and display data and information in "real time", offering an unprecedented opportunity to conduct a new form of knowledge discovery. Examples include artificial intelligence, machine learning, deep learning, scientific workflows and redefining what data actually is with the ability to study the kinds of data that are represented in the lower levels of Jennex's revised knowledge management pyramid (e.g., voice, image and text). With this shift has also come rethinking from scholars within the contributing disciplines, such as William S. Cleveland's "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" (2001) and Thomas H. Davenport and Jeanne Harris's "Competing on Analytics" (2007). These authors, and others, view the emerging discipline of data science as a transformed and new field of science that has extended beyond the walls of the academy into industry all the way to the more granular level of curiosity fueled by societal connectivity.

Multi-Modal Paradigms for Scientific Discovery and Knowledge Production

It has traditionally been assumed that the strength of an academic discipline resides in its purity, integration and in its distinctness (GlobalHigherEd, 2017). However, scientific discovery and knowledge expansion is less likely to occur in the well-established and distinct "center" of an academic discipline and more likely to occur on the less-established edges. Consider the development of relatively nascent, but increasingly accepted entrants into the academy such as biochemistry, astrobiology and environmental science. These new and emerging disciplines evolved at the intersection of fringes of their "parent disciplines" that were created through new developments and changing societal needs – much like species evolving to take advantage of a changing climate or ecosystem.

Most papers, essays and books on the birth of disciplines invariably begin with the current operational definition of the phenomenon (said new discipline), quickly followed by a history of the development and transformational evolution of the field to its current state. While helpful from a historical perspective, this type of inquiry does not lend context to the process of creation for new disciplines, the factors that have led to them or provide any indication regarding the discipline's trajectory.

These new disciplines are what Gibbons et al. foretold as manifestations of a new mode of knowledge production and scientific inquiry (Baber & Gibbons, 1995). The organizing principal of this evolved mode of knowledge production is that it affects not only what knowledge is produced but also how it is produced, the context in which is it pursued, the way it is organized, the reward systems it utilizes and the mechanisms that control quality.

Informed by Gibbons' framework, the evolution of data science as an academic discipline can be contextualized using the lens of academia with an eye towards a sustained and formal contribution to the revised knowledge management pyramid (Jennex, 2017).

Any discussion of data science as an evolving academic discipline must be caveated with an acknowledgement that there are many in the related sciences who do not recognize anything here more than an extension of their community's contributions. In this discussion, a nascent scientific discipline is posited for consideration – but not one that is in anyway displacing traditional and well-established disciplines such as statistics or computer science. To that point, Gibbons et al. emphasize that an evolved mode of knowledge production will emerge alongside and extend, rather than replace, the traditional, established mode.

The differences amongst defining attributes between the evolved mode of knowledge production – referred to by Gibbons as "Mode 2" – and the traditional mode of knowledge production – referred to as "Mode 1" are summarized in Table 1.

The authors posit that the evolution of data science can be explained using Gibbons' framework of a "Mode 2" discipline of knowledge production; without the attributes present in Mode 2, data science would not/could not become a unique academic discipline. Each of these modal attributes and the impact on the evolution of data science are explored below.

Knowledge Creation

The relevant contrast between the two modes is between problem solving organized around the codes of practice relevant to an established discipline and problem solving organized around a particular application.

Data science is generally accepted as an application-driven discipline that inherited its academic DNA from two more theoretically-driven disciplines – statistics and computer science, with both emanating from applied mathematics. In evolutionary biology, the term "heterosis" refers to the improved fitness of a hybrid offspring. For academic disciplines, this refers to the phenomenon that knowledge production drawing from research with input from multiple disciplines is of superior quality to research that limits its scope to a single academic silo (Cohen & Lloyd, 2014).

The datafication of the economy, as referenced above, created unforeseen data-centric challenges and opportunities, pushing both statistics and computer science out of their centers and into their respective fringes. Much of the traditional core of statistics, which has been established to accommodate data that is small, structured and static becomes less relevant where problems are defined by data that is large, unstructured and in-motion. Money and Cohen (2018) similarly recognize the shortcomings of traditional statistics to the Big Data environment:

…the present analysis techniques are rooted in statistical analysis, and significance tests that are irrelevant due to the population sampling and subsequent generalization to an entire population. In the contrasting Big Data analysis, Big Data sets or streams are not samples, they are massive and represent the majority of or a full population. The concept of statistical significance is not particularly relevant to Big Data.

Similarly, while the core of computer science enables the efficient capture and storage of massive amounts of structured and unstructured data, the discipline is less able to accommodate the need to translate that data into information through modeling, classification and visualization.

The datafication of all sectors of the economy – healthcare, manufacturing, finance, and retail – has been the source of these ongoing data-centric challenges and opportunities. These applications were the catalysts to fuse statistics and computer science at their fringes, creating an academic heterosis in response to the emergence of a new problem set, for which each siloed discipline was ill-equipped to address.

Table 1. Gibbons four modal attributes of knowledge production

Attribute	Mode 1(Traditional)	Mode 2 (Evolved)
Knowledge Creation	Primarily theoretically driven	Primarily application driven
Span of Engagement	Single discipline	Transdisciplinary
Diversity of Engagement	Primarily academic	Academic/Private sector collaboration
Quality Control	Centralized	Open accountability and reflexive

The greatest "proof" of a discipline earning academic citizenship as a mode of knowledge production is the advent of peer-review publication outlets. There are two notable aspects to the academic journals dedicated to data science – both related to the core of data science being a discipline in application.

The first aspect is that data science is emerging as a "horizontal" discipline, which crosses multiple "vertical" domain applications.

Regardless of the journal, the emphasis is on the application of data science to the domain rather than on the domain in question. Importantly, where there were theoretical articles, the emphasis was on the fused fringes of new algorithms or incremental improvements to emerging data science methods like machine learning or deep learning.

The second aspect of these peer-reviewed articles is that about a quarter of the authors were not affiliated with a university. A particularly relevant example of this emerging trend is one of the citations in this paper – Money and Cohen (2018) – is authored by an academic (Money) and a practitioner (Cohen). This creates interesting challenges to the foundations of how evidence has traditionally been explored, vetted and ultimately adopted as premise by both the academy and society – an issue explored later in this paper.

Span of Engagement

The evolved mode of knowledge production is grounded in a transdisciplinary approach to research – as opposed to a uni-disciplinary approach. It is worth noting here that there are conceptual differences across the terms "multi-discipline", "inter-discipline" and "trans-discipline" which are more than just semantic. Multidisciplinarity draws on knowledge from different disciplines but stays within their boundaries. Interdisciplinarity analyzes, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole. Transdisciplinarity integrates sciences in context and application, and transcends their traditional boundaries (Choi & Pak, 2006). New knowledge produced in a transdisciplinary context likely does not retro-actively fit into any one of the disciplines that contributed to the solution. The concept of transdisciplines was crystalized by Stichweh in 2001:

Transdisciplines by their nature have the characteristic of bringing together fields that otherwise appear to have little in common, thereby helping to re-unify science as it was before science was cast into separate silos of knowledge and research traditions in the 19th century.

Table 2. A sampling of academic journals with "data science" or "Big Data" in the title

Journal	2017 Articles	Sample of Areas of Application
Data Science Journal (Datascience.codata.org)	17	Nuclear Power, Insurance, Climate Change, Federal Data Policy, Sustainability, Social Science
International Journal of Data Science and Analytics (Link.springer.com)	61	Minority Discrimination, ADHD, MRIs, Recommender Systems, Cell Phones in Smart Cities, Social Network Analysis
Big Data (Liebertpub.com)	18	Propaganda and Politics, Profit-Driven Analysis, Social and Technical Tradeoffs in Big Data, Minority Discrimination, Education
Journal of Data Science (Jds-online.com)	20	Credit Scoring, Anemia in Married Females, European Football, Chemotherapy, Clinical Trials
International Journal of Data Science (Inderscience.com)	9	Supply Chain Management, Tuberculosis Screening, Minority Discrimination, Financial Distress Detection

In an academic environment, transdiscipline knowledge production will typically occur outside of the traditional discipline-oriented school or department. While a professor of biology and a professor of chemistry may co-author a research paper (an example of multidisciplinarity), a stream of biochemistry research which engages multiple faculty and graduate students is better accommodated through a separate university unit like a center or an institute, which can also facilitate collaboration with entities engaged in similar research outside the university. In addition, over time, these biochemists develop their own distinct theoretical structures, methods and outlets of dissemination of their research, which traditional biologists or traditional chemists would likely not value at the same level as their communities' traditional outlets.

Data science is following a similar path of disciplinary evolution. A sample of masters-level programs in data science at 45 universities across the United States revealed that 21 were housed outside of a siloed academic college and were housed in a "Mode 2" knowledge production unit.

Similarly, a sample of Ph.D.-level programs in data science, revealed that five out of 13 were housed in a "Mode 2" knowledge production unit.

The non-siloed units which house these programs are characterized by explicit statements of transdisciplinarity, allowing for short term appointments and fellowships from faculty and external contributors to defined research initiatives. These units also become a source of dissemination of research and knowledge production – in some cases bypassing the peer review process altogether and creating a searchable repository for research products (e.g., Digital Commons).

Diversity of Engagement

Mode 2 knowledge production is diverse in terms of the skills and experience that contributors bring to it. This is a function of the applied orientation. Gibbons' framework identifies that Mode 2 is marked by an increase in the types of units that collaborate to produce knowledge; no longer only universities but non-university research centers, government agencies, industrial laboratories, consultancies and private sector organizations. Weinberger (2012) too suggests that the emergence of Big Data is changing the shape and evolution of knowledge management to less structured "networks" both within and across organizations. In such environments, patterns of funding exhibit a similar

Figure 1. U.S. masters level programs in data science by university location[1]

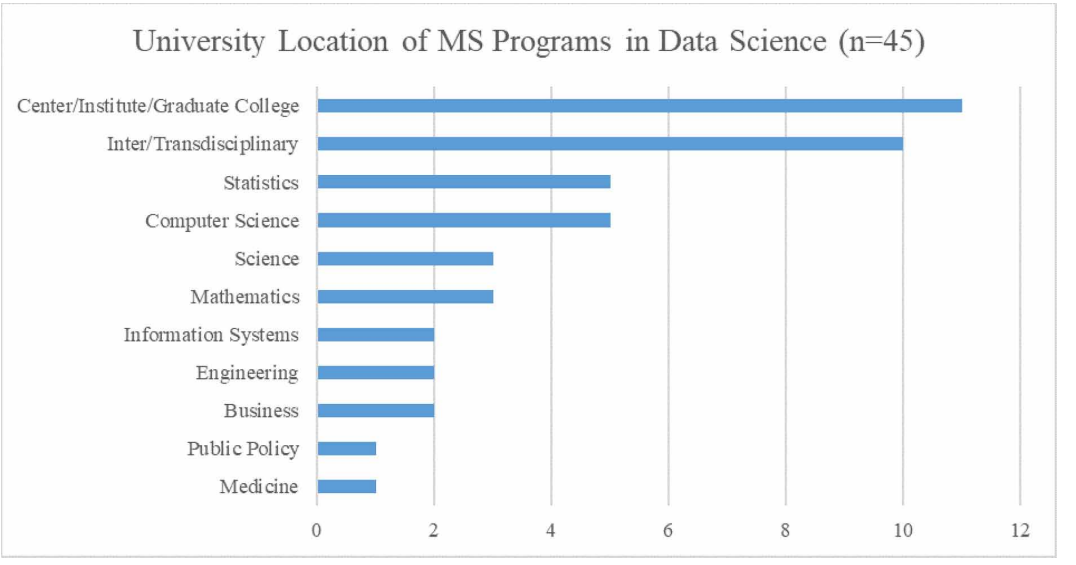
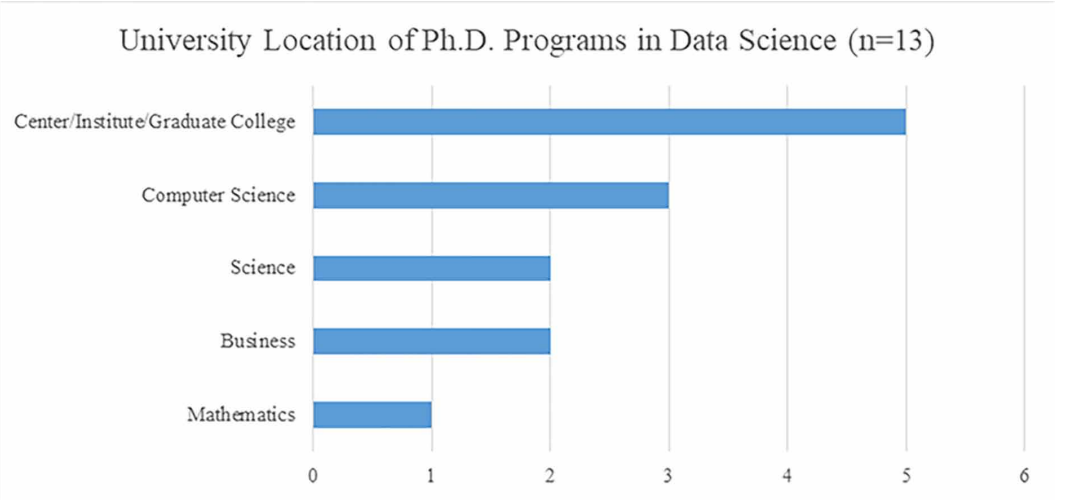


Figure 2. U.S. Ph.D. level programs in data science by university location[1]



diversity, being assembled from a variety of organizations with a diverse range of expectations and requirements, but all aligned with application.

As evidenced above, the number of centers, institutes and other transdisciplinary units with an emphasis on research and the dissemination of knowledge in data science is growing. However, while some of these units exist as transdisciplinary within the bounds of the university, a scan of most of these units reveals a roster of diverse participants. For example, the National Consortium for data science (NCDS) housed at the University of North Carolina, includes fellows from North Carolina State University and Drexel University, but also thought leaders in data science from IBM, Deloitte and the Environmental Protection Agency (EPA). The NCDS provides the following description of their organization:

The National Consortium For Data Science (NCDS) is a collaboration of leaders in academia, industry and government formed to address the data challenges of the 21st century. The NCDS was founded as a mechanism to help the U.S. take advantage of the ever increasing flow of digital data in ways that result in new jobs and industries, new advances in healthcare, transformative discoveries in science, and competitive advantages for U.S. industry.

A second example is the MIT-IBM Watson AI Lab. The Lab is a joint venture between MIT and IBM, housed on the MIT campus. The venture brings together MIT students and faculty with IBM researchers:

The MIT-IBM Watson AI Lab is one of the largest long-term university-industry AI collaborations to date. Our Lab is a place where scientists, professors and students collaborate to drive the frontiers of AI...the lab's scientists and engineers will focus on fundamental scientific breakthroughs, publish their results, and help guide the development of AI. A distinct objective of the lab is also to encourage MIT faculty and students to launch new companies that will focus on commercializing AI inventions and technologies that are developed at the lab.

There is another reason these transdisciplinary or multi-stakeholder entities will grow in importance. As Money and Cohen (2018) identify, the very work of managing networks of streaming

data ecosystems – the IoT layer in Jennex's revised pyramid – that will fuel future AI such as smart commerce, smart cities and more will be dependent on a continual set of data activities which are intensive but for which the benefits accrue downstream beyond any one of the data purveyors. These include data quality, statistical and algorithmic validity, data integrity and managing architecture and its development.

The types of diverse needs of thought, skills and orientation are becoming representative of the research, discovery and application products of data science. This is true, in part, because of the pace of technology, the size of the data and the multidimensional aspects of data-centric challenges, with no one contributor having the skills, knowledge or funding to address on a uni-lateral basis. It is also indicative of an employment trend that has been growing for decades – more than 50% of individuals who earn a Ph.D. in a scientific discipline work in the for-profit sector (US News and World Report, 2016). This has contributed to a more permeable membrane between the academic and private sectors, and has expanded the breadth of collaboration and funding in scientific research, scholarship and knowledge production.

Quality Control

The definition of "quality" research in traditional (Mode 1) knowledge production has been well established for decades. This has been particularly true in the sciences. Traditionally, the process of establishing agreed upon "evidence" was a prolonged process largely controlled by universities and/or science-based government agencies, or what is roughly termed "the Academy". It is a term dating back to the teachings of Plato, but which is more currently defined as "a body of established opinion widely accepted as authoritative in a particular field". Within the academy, research is conducted according to defined methods (taught by the academy), building on prior research, and utilizing the most accepted design and analytical methods. The proposals for such research were often funded by agencies, but also foundations, and the results were often compiled and published via the peer review process in scientific journals for the academic community to discuss. The peer review process consists of journals of varying reach and impact with editorial boards of reviewers who give unbiased feedback into a paper's merit, method and findings. These boards are also populated from members of the academy.

Membership in the academy, at minimum, typically requires that one have a terminal degree from an accredited university. The process has been criticized over the years as having a number of inherent limitations, such as professional competition, disagreement on foundational beliefs, and more currently, as being subject to outside influences due to the proliferation of online and for-profit "journals" that do not subscribe to original tenets of the process or that allow for gaming of the process (Kassirer & Campion, 1994; Ferguson & Oransky, 2014).

Socially distributed science and knowledge presents a new challenge where evidence verification is not necessarily expertly derived via the peer review process but through impact and utilization measures such as code being used, libraries accessed, and verified impact of methods in real application. Papers still exist, but in open source with open comment.

This, of course, gives rise to questions of the quality control of such findings and methods for which the academy was aptly designed to address. For example, what is the potential for corruption of the process; were the methods transparent and reproducible (and does it matter if the outcome is verifiable); and, what are the agreed upon standards for the process?

Generations of early career academics have had to pursue a single path to recognition and/or academic tenure through the process of peer review. In this context, the process of quality control is maintained by those deemed competent by the academy to act as "peers". The influence of these intellectual gatekeepers to traditional knowledge production cannot be overstated, including professional control through peer review to define what problems and techniques are even deemed important. Non-traditional findings or approaches to new challenges may be dismissed because they are evaluated with rubrics from the "core" rather than from the "fringe".

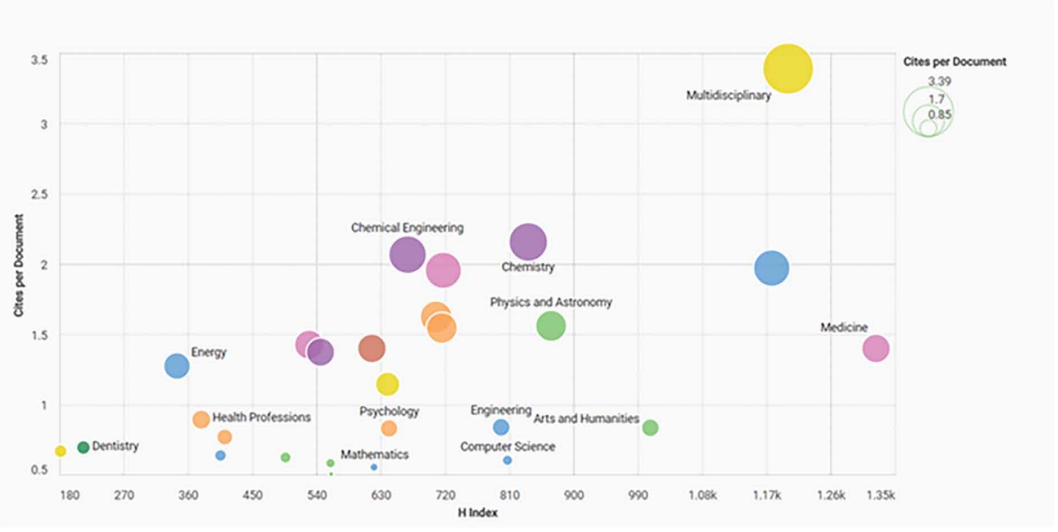
This traditional peer review process is one in which quality and control mutually re-enforce one another and can suffer from intellectual incest. In evolutionary terms, inbreeding occurs when an offspring has a parent pairing where considerable genetic material is shared, which leads to a decrease in species' fitness for survival. Single, highly siloed diciplines are subject to intellectual inbreeding, resulting in echo chambers that lead to low impact research (Cohen & Lloyd, 2014).

An examination of 2016-2017 published journal articles and conference proceedings indicates that multidisciplinary research generated more citations per article than any individual discipline and trails only the medical discipline in h-index score (Scimagojr.com).

In the evolved Mode of knowledge production, additional criteria are added as a function of the application orientation, where parties with a range of intellectual and practical interests will contribute to the evaluation of the definition of "quality". Since quality is no longer limited strictly to the review of academic "peers" and centralized control is now distributed in this mode there could be a concern of lower quality production that "falls through the cracks". However, because the control process is more broadly based and will experience multidimensional scrutiny from a new and expanded definition of "peers", an assumption of decrease in quality does not follow.

Knowledge production from data science is particularly subject to multidimensional scrutiny because of growing public awareness and concern about data-centric issues related to the environment, public health, and privacy. Any data-centric project that engages health care data will be subject to the Health Insurance Portability and Accountability Act (HIPAA), education data will be subject to the Family Education Rights and Privacy Act (FERPA) as well as an Institutional Review Board (IRB) and consumer financial data will be subject to a myriad of regulations including the Payment Card Industry Data Security Standard (PCI DSS), the Fair Credit Reporting Act, the Graham Leech Bliley Act, as well as the communications of the Consumer Financial Protection Bureau. An expanding number of public interest groups will likely express interest (and in some cases concern) related to the impacts of data-centric research. Those who work regularly with data have a built-in sensitivity to the impact, the perception and reception of their knowledge production – because they have to. This sensitivity has created a discovery environment that is more reflexive and more responsive to a broader audience than any earlier discipline; the need to be socially accountable effectively precludes the production of data science knowledge in Mode 1.

Figure 3. Impact of research by discipline



THE FUTURE OF DATA SCIENCE AS A MODE OF KNOWLEDGE PRODUCTION

Data science is evolving as a separate academic discipline in a Mode 2 framework for knowledge production. Specifically, the application-driven problem sets across all sectors of the economy have contributed to the fusing of computer science and statistics to form this new discipline, which is transdisciplinary, characterized by collaboration with the private sector and exhibits increased accountability and reflexivity.

However, one possible outcome of evolution is always extinction. Unlike creatures that can completely die out, knowledge – particularly in a digital age – never really completely disappears. Consider the knowledge related to replacing a typewriter ribbon or programming a VCR. However, academic disciplines are on a path to extinction when three things happen: (1) they consistently experience low student enrollment numbers because of a lack of demand for the skills and the knowledge produced, (2) little or no impactful research is being conducted or published, and (3), the tools and skills of that discipline are no longer thought useful to society.

While the immediate future of data science as a mode of knowledge production is bright, its long-term future is not ensured. Universities and the larger academy are faced with a new discipline that will require a new orientation towards knowledge production, which is currently modeled by data science. Any new orientation towards knowledge production comes with inherent challenges at multiple levels:

1. The data driven ecosystem will require collaborative and convergent models to support growing technological abilities and needs as data becomes faster, interconnected, larger and more integrated into virtually all aspects of human life. These will require thoughtful models of incentives to ensure those data and systems continue to produce and inform valuable knowledge outputs;
2. Research silos, while constructed for important reasons, may no longer reflect the intellectual landscape of society. (Gill, 2013; Gazzaniga, 1998) Silos further contribute to intellectual inbreeding and to research echo chambers. The future of high-impact, relevant research is in transdisciplinary communities. The process of promotion and tenure needs to embrace and reflect a more transdisciplinary orientation as well as more democratic methods of knowledge production and quality control of evidence creation;
3. The academy has traditionally developed and employed a number of research "specialists" where new research "generalists" may be needed that span areas of impact and implication to work across silos in order to examine highly integrated societal contexts. (Cohen & Lloyd, 2014) For example, medicine and health care in the United States has begun to focus on the socioeconomic, sociodemographic, and social determinants of health outcomes and spending beyond those drivers of cost and quality derived from health services utilization (Joynt, 2017);
4. Universities are typically slow to respond to societal or industry focused problems. New fields such as data science rely on both private and public contexts as well as on transdisciplinary partnerships, which require more flexibility and responsiveness (Borrego & Newswander, 2011);
5. University structures need to be nimbler and made to include more joint faculty focused on transdisciplinary work. Institutes, centers and possibly schools can accomplish this, however it requires rethinking the function and structure of both research and teaching as well as service and their relationship to the connections between home departments and transdisciplinary focal areas;
6. Increasingly more students who pursue a Ph.D. in a given field may never go into academia, however they may engage in meaningful impactful research. The membrane between academia and the private sector is becoming increasingly more permeable. The exponential pace of change and growth in technology and data complexity places added emphasis on universities to be a more active part of the division of labor that is knowledge creation. Not all industries can or should replicate the resources necessary to engage in knowledge

production as these are often duplicative across industries and would increase the cost of production. Universities can partner with organizations and institutions to fill this role. Thus, Ph.D. programs need to not just accommodate this new orientation, but embrace and leverage it to produce graduates who can contribute to meaningful research in larger, more impactful ways than just writing for academic journals.

This paper began with the question of whether data science is a new field of scientific study or rather a specialization within an existing discipline such as statistics or computer science. Informed by Gibbons et al.'s framework for evolved knowledge production, the authors posit that data science is emerging as a new transdisciplinary field and an evolved mode of knowledge production. The factors that have contributed to its (to date) success can be used to inform and reconsider the other currently siloed academic disciplines to meet the demands of a 21st century education, with particular emphasis on collaboration with the private sector and increased accountability and reflexivity.

REFERENCES

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.

Baber, Z., Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1995). The new production of knowledge: The dynamics of science and research in contemporary societies. *Contemporary Sociology*, *24*(6), 751. doi:10.2307/2076669

Bender, E. (2016). Challenges: Crowdsourced solutions. *Nature*, *533*(7602), S62–S64. doi:10.1038/533S62a PMID:27167394

Blanco, S. (2016). CEO Mark Fields says Ford will become a data company. Retrieved from http://www.autoblog.com/2016/01/06/ceo-mark-fields-ford-data-company/

Borrego, M., & Newswander, L. K. (2011). Analysis of interdisciplinary faculty job postings by institutional type, rank, and discipline. *Journal of the Professoriate*, *5*(2).

Choi, B. C., & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: Definitions, objectives, and evidence of effectiveness. *Medecine Clinique et Experimentale [Clinical and Investigative Medicine]*, *29*(6), 351–364. PMID:17330451

Cleveland, W. S. (2001). Data Science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. doi:10.1111/j.1751-5823.2001.tb00477.x

Cohen, E., & Lloyd, S. (2014). Disciplinary evolution and the rise of the transdiscipline. *Informing Science: The International Journal of an Emerging Transdiscipline*, *17*, 189–215. doi:10.28945/2045

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., & Players, F. et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760. doi:10.1038/nature09304 PMID:20686574

Datascience.codata.org. (2017). Data Science Journal. Retrieved from http://datascience.codata.org/

Datascienceconsortium.org. (2017). The National Consortium for Data Science. Retrieved from http://datascienceconsortium.org/

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics*. Boston, MA: Harvard Business Review Press.

Dee, S. (2015). How does capital one differentiate itself in the card industry? *Forbes*. Retrieved from https://www.forbes.com/sites/greatspeculations/2015/09/11/how-does-capital-one-differentiate-itself-in-the-card-industry/#407c5bff3cd

Drucker, P. (1995). The information executives truly need. *Harvard Business Review*, *73*(1), 54–62.

Dumont, J. (2016). The data revolution. *Techcrunch*. Retrieved from https://techcrunch.com/2016/10/06/the-data-revolution/

Ferguson, C., Marcus, A., & Oransky, I. (2014). Publishing: The peer-review scam. *Nature*, *515*(7528), 480–482. doi:10.1038/515480a PMID:25428481

Gardner, E. (2015). Look beyond academia to find jobs with a science Ph.D. *USNews*. Retrieved from https://www.usnews.com/education/best-graduate-schools/top-science-schools/articles/2015/03/30/look-beyond-academia-to-find-jobs-with-a-science-phd

Gazzaniga, M. S. (1998). How to Change the University. *Science*, *282*(5387), 237–237. doi:10.1126/science.282.5387.237

Gill, T. G. (2013). Culture, complexity, and informing: How shared beliefs can enhance our search for fit-ness *Informing Science*. *The International Journal of an Emerging Transdiscipline*, *16*, 71–98. doi:10.28945/1778

Global Higher Ed. (2007) *The challenges of creating hybrid disciplines and careers: a view from Sweden*. Retrieved from https://globalhighered.wordpress.com/2007/12/03/the-challenges-of-creating-hybrid-disciplines-and-careers-a-view-from-sweden/

Hand, E. (2010). Citizen science: People power. *Nature*, *466*(7307), 685–687. doi:10.1038/466685a PMID:20686547

Inderscience.com. (2017). *International Journal of data science (IJDS) - Interscience Publishers.* Retrieved from http://www.inderscience.com/jhome.php?jcode=ijds

James Manyika, M. C. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey*. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Jds-online.com. (2017). Journal of data science. Retrieved from http://www.jds-online.com/

Jennex, M. (2017). Big Data, the Internet of Things, and the Revised Knowledge Pyramid. *The Data Base for Advances in Information Systems*, *48*(4), 69–79. doi:10.1145/3158421.3158427

Jennex, M., & Bartczak, S. E. (2013). A Revised Knowledge Pyramid. *International Journal of Knowledge Management*, *9*(3), 19–30. doi:10.4018/ijkm.2013070102

Joynt, K. E., De Lew, N., Sheingold, S. H., Conway, P. H., Goodrich, K., & Epstein, A. M. (2017). Should medicare value-based purchasing take social risk into account? *The New England Journal of Medicine*, *376*(6), 510–513. doi:10.1056/NEJMp1616278 PMID:28029802

Kassirer, J. P., & Campion, E. W. (1994). Peer Review: Cured and understudied, but indispensable. *Journal of the American Medical Association*, *272*(2), 96–97. doi:10.1001/jama.1994.03520020022005 PMID:8015140

Kuhn, T. S., & Hawkins, D. (1963). The structure of scientific revolutions. *American Journal of Physics*, *31*(7), 554–555. doi:10.1119/1.1969660

Liebertpub.com. (2017). *Big Data* Retrieved from http://www.liebertpub.com/big

Link.springer.com. (2017). *International Journal of data science and Analytics - Springer* Retrieved from https://link.springer.com/journal/41060

Liyakasa, K. (2015). Alibaba CEO: 'We're not just an e-commerce company – We're a data company.' Retrieved from https://adexchanger.com/ecommerce-2/alibaba-ceo-were-not-just-an-ecommerce-company-were-a-data-company/

Madden, S. (2012). From Databases to Big Data. *IEEE Internet Computing*, *16*(3), 4–6. doi:10.1109/MIC.2012.50

Money, W. H., & Cohen, S. J. (2018). Our Knowledge Management Hubble May Need Glasses: Designing for Unknown Real-Time Big Data System Faults. *International Journal of Knowledge Management*, *14*(1), 30–50. doi:10.4018/IJKM.2018010103

Naur, P. (1992). *Computing: A Human Activity -- Selected Writings From 1951 To 1990*. Addison-Wesley, New York: ACM Press.

Netflixprize.com. (2017). Netflix Prize: Home. Retrieved from http://www.netflixprize.com/

Press, G. (2013). A very short history of data science. *Forbes*. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#2fed852455cf

Rijmenam, M. (2015). A short history of big data. Retrieved from https://datafloq.com/read/big-data-history/239

Scimagojr.com. (2017). Scimago Journal and Country Rank. Retrieved from http://www.scimagojr.com

Segal, I. (1962). Scientific discovery and the rate of invention. In The rate and direction of inventive activity: economic and social factors (pp. 441-458). Princeton, NJ: Princeton University Press.

Spender, J.-C., & Scherer, A. G. (2007). The Philosophical Foundations of Knowledge Management: Editors' Introduction. *Organization*, *4*(1), 5–28. doi:10.1177/1350508407071858

Stichweh, R. (2001). History of Scientific Disciplines. In J. Wright (Ed.), *History of International Encyclopedia of the Social and Behavioral Sciences* (pp. 287–290). Elsevier Science Direct. doi:10.1016/B978-0-08-097086-8.03048-8

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, *1*(2), 85–99. doi:10.1089/big.2012.0002 PMID:27442063

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, *33*(1), 1–67. doi:10.1214/aoms/1177704711

ENDNOTES

[1] Graduate programs in the United States with "data science" in the title