HHS Public Access

Author manuscript

Int J Data Sci Anal. Author manuscript; available in PMC 2020 March 01.

Published in final edited form as:

Int J Data Sci Anal. 2019 March ; 7(2): 81–86. doi:10.1007/s41060-018-0138-6.

Quant Data Science meets Dexterous Artistry

Ivo D. Dinov

Statistics Online Computational Resource, Department of Health Behavior and Biological Sciences, Department of Computational Medicine and Bioinformatics, Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Data Science is a bridge discipline connecting fundamental science, applied disciplines, and the arts. The demand for novel data science methods is well established. However, there is much less agreement on the core aspects of representation, modeling, and analytics that involve huge and heterogeneous datasets. The scientific community needs to build consensus about data science education and training curricula, including the necessary entry matriculation prerequisites and the expected learning competency outcomes needed to tackle complex big data challenges. To meet the rapidly increasing demand for effective evidence-based practice and data analytic methods, research teams, funding agencies, academic institutions, politicians, and industry leaders should embrace innovation, promote high-risk projects, join forces to expand the technological capacity, and enhance the workforce skills.

Common natural laws explain well-known dichotomies like "*a picture is worth a thousand words*" vs. "*a word may be worth a thousand pictures*" (1), "Taylor-series expansion" vs. "Fourier-series expansion" of continuous functions (2), or "Banach spaces" vs. "their algebraic duals" (3). Human expression of physical experiences and our reflection on abstract ideas cover the continuum – from uttering musical tones, painting canvases, and art sculpting, to conceiving fundamental geometric structures, interpreting non-Euclidian topological objects, identifying mathematical patterns, and inventing data analytics that explain complex environments or elucidate effective messaging and communication. A very similar polarity governs the Big Data Science theory and practice – making sense of enormous amounts of heterogeneous, multi-source, multi-scale, incomplete and incongruent data requires rigorous foundational training paired with dexterous artistic skills. As proxies of complex natural phenomena, Big Data provide a unique view into the intrinsic process organization, enabling information extraction, reinforced learning, model-free inference, and deep understanding of the underlying systems.

There is a clear evidence of the growing synergies between the four fundamental discovery paradigms – experimental, theoretical, computational, and data sciences (4). Applying innovative data science techniques to tackle challenging computational problems requires substantial investment in both – hard and soft skills. It is commonly acknowledged that Data

Corresponding author: Ivo D. Dinov, PhD, Professor and Director, Statistics Online Computational Resource, Department of Health Behavior and Biological Sciences, University of Michigan, 426 N. Ingalls Str. Ann Arbor, MI 48109, statistics@umich.edu, www.umich.edu/~dinov, Phone: 734-764-5557.

Science is an extremely transdisciplinary field. A less recognized characteristic of Data Science is its intrinsic dual reliance on quantitative basic sciences techniques as well as qualitative Artistry (5). The distinct, yet equally important, contributions of fundamental scientific principles and artistic vision scale opposite slopes of the majestic Data Science mountain scape. Blending human-machine interfaces and data-driven inference is required to improve the long-term odds of survival and increase the impact of future decision support systems that rely on large amounts of complex empirical evidence. This has nascent and direct implications on public-private partnerships, research funding priorities, and team-based quantitative-qualitative education.

The history of data science commenced at least half a century ago (6). A number of efforts from diverse computational and data science communities are working to formulate the core methodological principles for conceptualizing, representing, modeling and understanding of complex processes that are observed as large, heterogeneous and multisource information streams. A recent definition of complex data science problems, *X-complexities*, based on ubiquitous intelligence, *X-intelligence*, illustrates an unconventional way to represent computational processes and data problem-solving systems (7, 8). This holistic approach to data science innovation, training and practice relies on quantitative data literacy coupled with domain, network and behavioral, organizational, social, and environmental intelligence. Contemporary scientific discovery demands highly complementary skills and abilities to develop and effectively cooperate with others on data analytics that cover the continuum from data aggregation, harmonization, processing, descriptive analyses, and modeling, to simple, advanced, deep, predictive, and prescriptive analytics (7). A paradigm shift in scientific discovery is taking shape that fuels the emergence of data science innovation and transformative analytics. This new paradigm connects disparate scientific domains, provides incentives for interoperable technologies, and promotes continuous development, scalable and effective inference, and high-throughput reproducible analytics. This is also paired with the exponential increase of the critical development gaps between data potential and state-of-the-art capabilities (8). Data science pitfalls include equating it disciplinarily to one specific field (e.g., statistics), misinterpreting duality of data size and complexity, infrastructure and methodology, and interpreting “Big Data” phenometrically as a census, i.e., a window to the entire natural process, or population. The evolution of the Internet over the past 40 years provides striking parallels for the expected growth of data science. Developing capacity for complex data science thinking and corresponding methodologies will provide enormous opportunities for breakthrough research, technological innovation, social and economic progress (9).

Intersection of artistic humanity and quantitative science

The turn of the 20th century represented the golden age of “theoretical sciences”. This second scientific pillar greatly relied on highly individualized efforts. While working out the theory of General Relativity, Albert Einstein drew parallels between the first two scientific paradigms – experimental and theoretical sciences. He connected the musical “harmony of spheres” and symmetries of space-time. In this process, Einstein frequently bounced between pen-and-paper, describing the mathematical foundations, and violin-and-bow, playing Mozart and other classical composers. In the 21st century, innovative Data Science

teams will follow suit by connecting the first four scientific paradigms – experimental, theoretical, computational, and data sciences. Balanced networks of basic, exploratory, humanitarian, and quantitative experts will provide the necessary competence, dexterous skills, artistic abilities, theoretical rigor, and technological talents that are jointly critical for data-driven discovery and advanced scientific progress. The evolution of complex living organisms required the rise of network systems. Similarly, expansion of cooperation, deep specialization, uni-disciplinary approaches, and closed-scientific activities will give way to team-based transdisciplinary *quant-qual* scientific innovation. Yet, the balance between deep scientific knowledge and broad dexterous skills will need to be maintained (10).

Big Data Fundamentals

A 2001 action plan, that outlined the scope, challenges and approaches to develop the field of data science, advocated the expansion of mostly technical statistical methods (11). Specifically, this strategy proposed a data science curriculum including (1) multidisciplinary investigations and data analysis collaborations (25%), (2) models and methods for estimation and distribution-based probabilistic inference (20%); (3) theoretical foundations of data science supporting data computing and mathematical investigations, and evaluation (20%); (4) pedagogical curriculum focus from elementary school to postgraduate school and continuing education (15%).

The theoretical underpinnings of data science have not been developed, yet. To the point, there are no broadly accepted strategies for canonical representation of general data as computable objects, the mathematical modeling principles are not fully formulated, the statistical analytic techniques rarely apply to large incongruent, incomplete and heterogeneous datasets, and each application demands ad-hoc protocols that are hard to adopt to new case-studies. Although the foundations of Big Data Science are not fully understood yet, there are parallel efforts to develop theoretical frameworks for representation, high-throughput analytics, and model-free inference on Big Data. For instance, Compressive Big Data Analytics (CBDA) is one approach utilizing core principles of distribution-free and model-agnostic methods for scientific inference (12). CBDA iterates (re)sampling and inference to derive posterior likelihoods, probability approximations, or parameter estimates, which can be used to assess algorithm reliability, determine result accuracy, or extract important features via controlled variable selection. Another approach involves mathematical tensors for modeling and representation of multi-aspect data (13). Tensor decompositions of Big Data provides an effective mechanism for unsupervised data learning, context understanding, and quantitative analysis.

Launching a Data Science Career

A common question asked by current and former students, as well as second-career trainees, is "*How can I get into data science?*" There are three components that provide necessary, but not sufficient, conditions for success. These include *background, environment,* and *motivation*. The former condition demands sufficient *a priori* exposure to basic science and quantitative training principles. This includes some programming skills as well as background in calculus, linear algebra, and/or probability and statistics. The environmental

component stipulates the existence of an immersive ecosystem that provides data science drivers, challenges, educational and training curricula, and career opportunities. Background and environment are extrinsic elements, whereas *motivation* is an intrinsic trait that is extremely specific to the learner and cannot be cultured purely by exogenous forces.

At the University of Michigan, we recognize the diversity of student backgrounds, interests, motivations, expectations, and learning styles. Table 1 shows the prerequisites expected to increase the chances of successful completion of the Data Science and Predictive Analytics (DSPA) course (http://Predictive.Space).

In addition, MOOCs may be taken to build, refresh, or expand trainees' background knowledge, and to meet program prerequisites. Examples of remediation courses are provided in the DSPA self-assessment (pretest).

Desired Data Science Competencies

As data science is an extremely transdisciplinary field, outlining a set of specific knowledge outcomes or competencies is bound to be too narrow, too broad, or both. Yet, we have developed an outline suggesting skills and abilities that are likely to enhance the professional capacity, broaden the community participation, and improve data scientist productivity.

Trainees successfully completing the DSPA course are expected to have high-to-moderate competency in at least two of each of the three *competency areas* listed in Table 2.

Examples of two independent data science curricula are developed by Microsoft Research (14) and the University of Michigan Institute for Data Science (15).

Specific Actions

The broad community of data science methodologists, technologists, practitioners, consumers, stakeholders and policy makers needs to deeply engage in discussions, implementation strategies, and valuation protocols that cover some of the following items:

- Develop flexible curricula providing fundamental data literacy, defining core prerequisites and basic competencies, and outlining the specific skills necessary for data science trainees at all levels, from undergraduate students to postdoctoral fellows. For example, the University of Michigan Institute for Data Science has openly shared their graduate data science and predictive analytics curriculum, defined the prerequisite knowledge, and stated the expected outcome competencies for trainees going through their program (http://Predictive.Space).
- Introduce protocols to build transdisciplinary teams that involve basic and translational scientists, which can be sustained over time. The persistently high preterm birth rate in the US (1 out of 8) presents an example of an extremely challenging health and well-being challenge that requires a transdisciplinary solution. The Stanford University March of Dimes established a transdisciplinary Translational Research Center to address the problem of preterm birth in 2011. This initiative gathered a broad team of experts to investigate strategies, design

highly integrated and holistic datasets, and build the infrastructure to understand and preventing preterm births (16). Similar efforts will be necessary to tackle future computational and data science challenges by dissolving intellectual silo barriers and forming integrated problem-based team science, adopting of novel methods, technologies, and removing risk aversion.

- Recognize the value of open-science, including FAIR data sharing (17) and open-source code principles (18). Hording data, stashing information in silos, or keeping knowledge private will not prevent, but rather slow scientific progress, and delay the translation of discovery into practice.
- Tackle fundamental data science challenges, including Big Data representation, harmonization, aggregation, simulation, modeling, inference, and visualization (9, 19). This will require *truly* investigator initiated scholarly activities. Most of the latest requests for funding applications, program announcements, and other funding opportunities from federal and private foundations encourage broad participation, innovation, and high-risk scientific endeavors. However, in practice, many are explicitly risk averse, overvalue prior track-record, and discount bona fide innovation. Scientists serving on review panels can attest that rather than boards of innovation, some of these study sections tend to act as closed-door congressional committees where seniority, politics, and special interests are paramount (20).
- Explore new mechanisms for public-private-partnerships (PPP) that can be used to *support* data science projects that advance basic research, expand the computational infrastructure, enable high-throughout information extraction, and facilitate effective communication and dissemination. This also includes targeted funding (research, education and services), community building and networking, promotion of diversity, equity and inclusion. The impact of independent, asynchronous and decentralized investments and pursuits by individual organizations could be substantially enhanced by joint PPP activities leveraging capital, formation of complementary-expertise teams, cost-optimization, and translation of discoveries into useful products (21–23).
- Incentivize and promote continuous self-learning, informal education, and live-long learning. The continuity of the rapid technological advances and our ability to handle the increasing volume and complexity of information will demand. After all, life-long learning is highly associated with higher-paying jobs and higher impact on communities (24). Furthermore, just like continuous physical activity improves somatic performance, sustained cognitive stimulation tends to increase productivity and buffer against mental decline (25).
- Recognition that innovation is uncertain! Supporting and enticing high-risk projects will generate both failures and successes. Frequently, the effect and impact of successful discoveries are not materialized for decades. The roles of stochasticity and serendipity in scientific advancement should never be underestimated. Recall these important accidental examples (1) the chemist Leo Baekeland who in 1907 accidentally combined formaldehyde (organic precursor

compound) with phenol (waste product of coal tar) to produce the first non-conductive heat-resistant polymer (plastic), and (2) the microbiologist Alexander Fleming who in 1928 left for two weeks an staphylococcus bacterial experiment and later found that a mold contaminant inhibited the bacterial growth, effectively discovering the first antibiotic – penicillin. Given a chance, data science is bound to have many such serendipitous discoveries.

Future decision makers will need to be quantitatively literate experts is several domains; capable of articulating, communicating and continuously self-training. In addition, the successful data scientists will be accomplished artists that lead by example, effectively borrow techniques across scientific areas, employ tools developed by others, and ensure the interoperability of disparate resources. A historical perspective from one of the most accomplished artist/scholar, the polymath Leonardo da Vinci, may be insightful. In the early 1480's, da Vinci sent a job application to Ludovico Sforza, the Duke of Milan at the time. Da Vinci's 15-page application listed many of his scientific interests, military abilities, and architectural expertise; however his artistic genius the artist's most dexterous skill was only mentioned in the end of his letter. Da Vinci understood the important interplay between abstract scientific concepts, practical skills, and artistic potential. Contemporary data scientists need to take a similar holistic approach to their careers.

Skeptics will continue to question the scientific principles of exploratory science, model-free inference, and the artistry in Big Data discovery. Some may try to temper the enthusiasm for large-scale data-intense projects, e.g., interfacing between neuro-biological and computational sciences (26). Indeed, the foundations of Big Data science is not developed, yet. However, there are lots of parallels and (un)coordinated developments underway. Most are destined to fail, but some are certain to succeed. Innovation is by default risky, and we can't predict exactly which activities will bear fruit. Bleak views like "*expedient industrialization of neuroscience and the potential long-term importance of the personal, political, and commercial incentives driving it are causes for concern*" are overly pessimistic and more detrimental than constructive. We do need certain level of stochasticity in the process of formulating the core principles of Big Data science. The skeptics should be reminded that even Einstein missed a number of innovative discoveries. For instance, when he studied whether objects in space could divert light, he rejected the possibility that objects could influence each other no matter how far apart they were; "*Of course there is no hope of observing the Gravitational Lensing phenomenon directly*", he remarked (27).

It is also important to realize the benefits and limitations of human skills, information knowledge, and machine performance. Data scientists are "Jacks of all trades" that have to have breadth of knowledge, a wide range of skills, and agile ability to adopt in the rapidly evolving universe inundated with complex, dynamic, and multi-source data. Those that are doubtful of data scientists, or their scholarly achievements, should realize the knowledge depth, discovery potential, and practical efficiency of data-intense research. Big Data is a proxy of the underlying phenomena. When coupled with team-based, network-wide, open-scientific, and crows-sourcing approach to future scientific innovation, it forms an extremely potent mixture. Martin Clifford's 1677 notes on the poems of Dryden ("Jack of all Trades [Shoppes], that have variety, but nothing of value") and the 1770 Gentleman's Magazine

opinion (“Jack at all trades, is seldom good at any”) were extremely narrow minded and proven to be archaic in the modern industrialized societies. Similarly, data science industrialization, popularization, and direct applications will become indispensable common practice in most occupations, businesses, academic institutions, and government activities.

Globally, progress will likely depend on various geographic, political, and socio-economic factors. However, the momentum of this transformation from individual-based knowledge to collective network-oriented exploration and data-driven scientific discovery will be contingent on the immediate cooperative actions, appropriate resource allocations, innovative education strategies, and visionary policies. Although there is not yet a blueprint for an ideal data rainmaker, the ultimate data scientist will be a dedicated self-learner that is a skilled quant as well as a dexterous artisan.

References

1. Robinson L. 2016; Frank Crossley: A Man of Mettle. JOM. 68(7):1743.
2. Weinberger ED. 1991; Fourier and Taylor series on fitness landscapes. Biological Cybernetics. 65(5):321–330.
3. Ryan, RA. Introduction to tensor products of Banach spaces. Springer Science & Business Media; 2013.
4. Hey, T, Tansley, S, Tolle, KM. The fourth paradigm: data-intensive scientific discovery. Microsoft research; Redmond, WA: 2009.
5. Knuth DE. 1991; Theory and practice. Theoretical Computer Science. 90(1):1–15.
6. Donoho, D. 50 years of Data Science. Princeton NJ: Tukey Centennial Workshop; 2015.
7. Cao L. 2017; Data science: a comprehensive overview. ACM Computing Surveys (CSUR). 50(3):43.
8. Cao L. 2016; Data science: nature and pitfalls. IEEE Intelligent Systems. 31(5):66–75.
9. Cao L. 2017; Data science: Challenges and directions. Communications of the ACM. 60(8):59–68.
10. Schwartz MS, Sadler PM, Sonnert G, Tai RH. 2009; Depth versus breadth: How content coverage in high school science courses relates to later success in college science coursework. Science Education. 93(5):798–826.
11. Cleveland WS. 2001; Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review. 69(1):21–26.
12. Dinov ID. 2016; Volume and value of big healthcare data. Journal of Medical Statistics and Informatics. 4(1):1–7.
13. Papalexakis EE, Faloutsos C. 2016; Unsupervised Tensor Mining for Big Data Practitioners. Big data. 4(3):179–191. [PubMed: 27642720]
14. Hopcroft, J; Kannan, R. Foundations of data science. Microsoft Research. 2014. https://www.microsoft.com/en-us/research/publication/foundations-of-data-science/
15. Dinov, I. Data Science and Predictive Analytics: Biomedical and Health Applications using R. Springer International Publishing; 2018. 800
16. Stevenson DK, et al. 2012; Transdisciplinary translational science and the case of preterm birth. Journal Of Perinatology. 33:251. [PubMed: 23079774]
17. Wilkinson MD, et al. 2016The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. :3.
18. Marszalek RT, Flintoft L. 2016; Being open: our policy on source code. Genome biology. 17(1): 172. [PubMed: 27520968]
19. Dinov I. 2016; Methodological Challenges and Analytic Opportunities for Modeling and Interpreting Big Healthcare Data. GigaScience. 5(12):1–15.
20. Li D, Agha L. 2015; Big names or big ideas: Do peer-review panels select the best science proposals? Science. 348(6233):434–438. [PubMed: 25908820]

21. Villani E, Greco L, Phillips N. 2017Understanding Value Creation in Public-Private Partnerships: A Comparative Case Study. Journal of Management Studies.
22. Martin CL, Chun M. 2016; The BRAIN initiative: Building, strengthening, and sustaining. Neuron. 92(3):570–573. [PubMed: 27809996]
23. Brownson, RC, Colditz, GA, Proctor, EK. Dissemination and implementation research in health: translating science to practice. Oxford University Press; 2017.
24. Newman, BM, Newman, PR. Development through life: A psychosocial approach. Cengage Learning; 2017.
25. Fisher GG, Chaffee DS, Tetrick LE, Davalos DB, Potter GG. 2017; Cognitive functioning, aging, and work: A review and recommendations for research and practice. Journal of occupational health psychology. 22(3):314. [PubMed: 28358568]
26. Frégnac Y. 2017; Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? Science. 358(6362):470–477. [PubMed: 29074766]
27. Einstein A. 1936; Lens-like action of a star by the deviation of light in the gravitational field. Science. 84(2188):506–507. [PubMed: 17769014]

Table 1

Data science and predictive analytics (DSPA) prerequisites.

Prerequisites	Skills	Rationale
BS Degree or Equivalent	Quantitative training and coding skills as described below	The DS certificate is a graduate program requiring a minimum level of quantitative skill
Quantitative Training	Undergraduate calculus, linear algebra and introduction to probability and statistics	These are the entry level skills required for most upper-level undergraduate and graduate courses in the program
Coding Experience	Exposure to software development or programming on the job or in the classroom	Most DS practitioners need substantial experience with Java, C/C++, HTML5, Python, PHP, SQL/DB
Motivation	Significant interest and motivation to pursue quantitative data analytic applications	Dedication for prolonged and sustained immersion into hands-on and methodological research

Table 2

Data science and predictive analytics (DSPA) competencies.

Areas	Competency	Expectation	Notes
Algorithms & Applications	Tools	Working knowledge of basic software tools (command-line, GUI based, or web-services)	Familiarity with statistical programming languages, e.g., R, Python, SciKit, and database querying languages, e.g., SQL or NoSQL
	Algorithms	Knowledge of core principles of scientific computing, applications programming, API's, algorithm complexity, and data structures	Scientific application programming, efficient implementation of matrix linear algebra and graphics, elementary notions of computational complexity, and user-friendly interfaces
	Application Domain	Data analysis experience from at least one application area, either through coursework, internship, research project, etc.	Computational social sciences, health sciences, business and marketing, learning sciences, transportation sciences, engineering and physical sciences
Data Management	Data validation & visualization	Curation, Exploratory Data Analysis (EDA) and visualization	Data provenance, validation, visualization via histograms, Q-Q plots, scatterplots (ggplot, Dashboard, D3.js)
	Data wrangling	Skills for data normalization, data cleaning, data aggregation, and data harmonization/registration	Imperfections in big data include missing values, inconsistent string formatting, and heterogeneous file types
	Data infrastructure	Handling databases, web-services, Hadoop, multi-source data	Data structures, SOAP protocols, ontologies, XML, JSON, streaming
Analysis Methods	Statistical inference	Basic understanding of bias and variance, principles of (non)parametric statistical inference, and (linear) modeling	Biological variability vs. technological noise, parametric vs non-parametric (rank order statistics) procedures, point vs. interval estimation, hypothesis testing, regression
	Study design & diagnostics	Design of experiments, power calculations and sample sizing, strength of evidence, p-values, False Discovery Rates	Multistage testing, variance normalizing transforms, histogram equalization, goodness-of-fit tests, model overfitting
	Machine Learning	Dimensionality reduction, k-nearest neighbors, random forests, AdaBoost, kernelization, SVM, ensemble methods, CNN	Empirical risk minimization. Supervised, semi-supervised, and unsupervised learning. Transfer learning, active learning, reinforcement learning, multiview learning, instance learning