The professionalisation of data science

Michael A. Walker

Data Science Practice, Rose Technologies,
Denver, Colorado, USA
Email: michael.walker@datascienceassn.org

Abstract: Data science is establishing basic foundations to become a profession. Like the professionalisation of law and medicine in the past 100 years, the data science field is at the very beginning of becoming a profession – with competency standards, a code of professional conduct, specialised graduate-level curriculums, certification and licensure and self-regulation. All professions require highly specialised education and training, an ethical code, self-regulation by a professional association and certification and licensing. Data science should become a profession for the same reasons medicine and law became professions: each requires practitioners to have a specialised body of knowledge, a code of conduct and self-regulation by knowledgeable professionals to assure competency and protect the public. The data science community can follow a roadmap for how data science can be professionalised by reviewing the history of the medical and legal professions. Suggested is a seven-step process for the professionalisation of data science.

Keywords: profession; data science; licensure; certification; regulation; self-regulation; data science code of professional conduct; specialised knowledge; ethics; education.

Reference to this paper should be made as follows: Walker, M.A. (2015) 'The professionalisation of data science', *Int. J. Data Science*, Vol. 1, No. 1, pp.7–16.

Biographical notes: Michael A. Walker is the President of the Data Science Association, a non-profit professional association of over 4000 data scientists globally. He is the author of the 'Data Science Code of Professional Conduct' and is writing a book on data science strategy for business and government. His data science writings are widely publicised and cover topics such as data science strategy and techniques, algorithms, machine learning, business analytics, ethics and data technologies. He earned his Bachelor of Arts from the University of Colorado and Doctorate of Jurisprudence from Syracuse University.

1 Introduction

There is much confusion and debate about the definition of data science and the new rare breed of sexy bird called the data scientist. While salaries for business intelligence and data warehousing professionals are stagnating, data scientists are at the top of the pay scale according to a recent 2014 InformationWeek/Burtch Works Salary Survey (Table 1).

Table 1 Burtch Works 2014 Survey

Job title	*Median staff salary*	*Median mng. salary*
Bl/Analytics	$87,000	$110,000
Data integrating/Warehousing	$100,000	$120,000
Big Data Professionals*	$90,000	$145,000
Data Scientists*	$120,000	$160,000

Sources: InformationWeek Salary Survey 2014/Burtch Works*

The Data Science Association defines 'Data Science' as the scientific study of the creation, validation and transformation of data to create meaning, and the 'Data Scientist' as a professional who uses scientific methods to liberate and create meaning from raw data.

While these definitions may appear overbroad, think about the definitions of a lawyer or a physician. A lawyer is a legal professional who can help prevent or solve legal issues and a physician is a health professional who can help prevent or cure health issues. Like the professionalisation of law and medicine in the past 100 years, data science is at the very beginning of becoming a profession – with competency standards and a data science code of professional conduct.

Members of the data science community can gain a new perspective about how the data science field can be professionalised by reviewing the medical and legal professions.

2 Definition of profession

According to Wikipedia:

> "A profession is an occupation that requires extensive training and the study and mastery of specialised knowledge, and usually has a professional association, ethical code and process of certification or licensing."

Professionalisation is defined as a job that requires special education, training or skill (Merriam-Webster Dictionary). It may be described as the process that a trade or occupation transforms itself into a true profession that establishes membership qualifications and a professional association that proscribes norms of professional conduct (Bullock and Trombley, 1999). Educational standards, licensure, certification and accreditation attempt to assure the public of professional competency and create a degree of demarcation between qualified and unqualified practitioners. To become a real profession usually means obtaining the legal right to self-regulate.

The following are examples of occupations that transformed into professions:

- physicians
- lawyers
- dentists
- accountants
- architects.

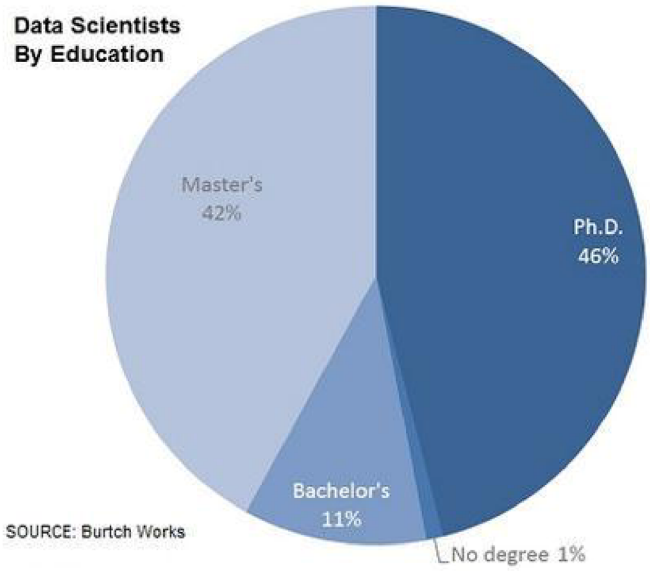
Physicians and lawyers were among the first to professionalise successfully. They became self-regulated pursuant to legal codification and could set standards for education, licensure, practice and behaviour. Professions like medicine and law have common attributes, including:

- core body of specialised knowledge
- associations to set training, education and ethical guidelines
- code of professional conduct
- commitment to continuing professional development
- self-regulation.

Different levels of education, skill development and ongoing accreditation are required for different professions. In the future, data science will be influenced by data science practitioners, outside organisations and both federal and state governments.

According to a recently released report by Burtch Works, 46% of data science professionals have a Doctorate, 42% a Master's Degree, 11% a Bachelor's Degree and 1% no degree (Figure 1).

Figure 1 Burtch works 2014 survey (see online version for colours)



3 Should data science become a profession?

Should data science become a profession? Should regulations be enacted governing the practice of data science and who should create those regulations?

Data science should become a profession for the same reasons the practice of medicine and law became professions: each requires special education, training and skill, each requires a code of conduct and the malpractice of each specialised domain may

cause severe negative consequences (i.e., bodily injury, death, incarceration, loss of money or property, privacy violations and bad public policy). Thus, each domain requires educational standards and a code of conduct to assure the public of competency.

At this time, the field of data science is overcrowded with unqualified practitioners. There is market confusion about the definition of data science and the skills required of a data scientist. Many garden variety data analysts have renamed their job titles as 'data scientists' to exploit market confusion and attempt to gain unwarranted status and salary increases. This is creating competition between the 'qualified' and the 'unqualified,' and hurting those who are truly qualified to practise. In addition, this is leading to harm being done to clients and employers. These problems lead to widespread demands to create specialised data science programs at the university undergraduate and graduate levels and establish a system of registration and minimum training requirements to practise data science.

While garden variety data analysts are abundant, today there is an acute shortage of qualified data scientists. Recent studies by McKinsey and others conclude that there will continue to be a shortage of real data scientists in the near future. In response, a number of university graduate degrees and online courses for data science and related advanced analytics have been created. While on balance a positive development, there still is a need for a data science accreditation organisation to establish educational guidelines and performance standards, and to create a clear demarcation between qualified and unqualified practitioners.

This laid the foundation for establishing the Data Science Association, an organisation built to develop proficiency in existing skills, develop new skills and set new conduct and performance standards. The Data Science Association was formed to advance data scientific disciplines, define and improve standards in data science education and establish a code of ethics – with the overall public goal of improving life, business and government using advanced data science techniques.

Evidence suggests the practice of data science requires special education, training and skill and misconduct is common in both the hard and soft sciences. Dr. John Ioannidis' 2005 paper 'Why Most Published Research Findings Are False' provides strong evidence of flawed science and misconduct among professional scientists. He analysed 49 of the most highly regarded research findings in medicine over the previous 13 years. He compared the

> "45 studies that claimed to have uncovered effective interventions with data from subsequent studies with larger sample sizes: 7 (16%) of the studies were contradicted, 7 (16%) of the effects were smaller than in the initial study and 31 (68%) of the studies remained either unchallenged or the findings could not be replicated."

The Atlantic magazine summarised Ioannidis' findings:

> "His model predicted, in different fields of medical research, rates of wrongness roughly corresponding to the observed rates at which findings were later convincingly refuted: 80% of non-randomised studies (by far the most common type) turns out to be wrong, as do 25% of supposedly gold-standard randomised trials, and as much as 10% of the platinum-standard large randomised trials… 'You can question some of the details of John's calculations, but it is hard to argue that the essential ideas are not absolutely correct,' says Doug Altman, an Oxford University researcher who directs the Centre for Statistics in Medicine."

While Ioannidis focuses on medical research, his findings likely apply to other scientific disciplines, especially in the soft sciences that use data science techniques. For example, many academic or research scientists run thousands of computer simulations where all fail to confirm or verify the hypothesis. Then, they tweak the data, assumptions or models until confirmatory evidence appears to confirm the hypothesis. They proceed to publish the one successful result without mentioning contradictory evidence or the thousands of failures! This is unethical, may be fraudulent in some cases and certainly produces flawed science where a significant majority of results cannot be replicated. This has created a loss or confidence and credibility for science by the public and policy-makers that has serious consequences for our future.

Furthermore, it is well to remember that raw datasets – both large and small – are not objective. They are selected, collected, filtered, structured and analysed by human design. What was measured, in what manner, with what devices and to what purpose? What was not measured and why? Was only low-hanging fruit measured because the important could not be measured? What was the quality of the data? It is disheartening that many scientists refuse to disclose raw data and data collection methods.

Humans then interpret meaning from data in different ways. Data scientists can be shown the same sets of data and reasonably come to different conclusions. Naked and hidden biases in selecting, collecting, structuring and analysing data present serious risks. How we decide to slice and dice data and what elements to emphasise or ignore influences the types and quality of measurements.

The danger for professional data science practitioners is providing clients and employers with flawed data science results leading to bad business and policy decisions that have the potential to damage both individuals and institutions. Data scientists must learn from the academic and research scientists and proactively avoid confirmation bias or data science risks loss of credibility.

Therefore, it is critical for data scientists to follow a code of conduct and develop data science processes to guard against misconduct and to ensure the public of professional competency to uphold the reputation and maintain credibility of data science.

Moreover, at this time many garden variety business or data analysts are attempting to capitalise on the hot job title of ‘data scientist’ exploiting market confusion over the definition of data science. Evidence suggests that in many organisations garden variety data analysts with no or little scientific training, deep analytical training or experience have renamed themselves ‘data scientists’ and are falling into a myriad of ‘bad science’ traps (e.g., cherry picking, data selection bias, confirmation bias, narrative fallacy, Texas sharpshooter fallacy, cognitive bias – among others). Many commit misconduct causing their clients or employers real damage. Like confusing signal and noise can lead to tragedy, confusing professional data scientists with garden variety business and data analysts can lead an organisation to disaster.

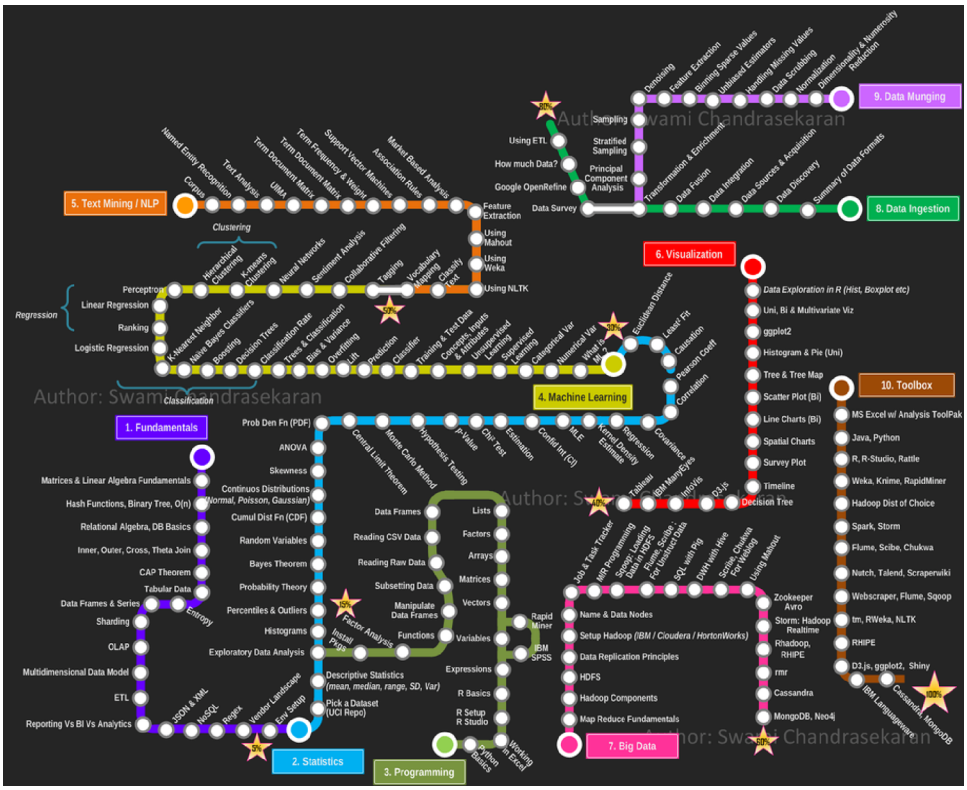
For the above-listed reasons, data science must become a profession.

4 Specialisation of data science

The practice of law has about 24 different practice areas. The practice of medicine has approximately 17 specialties. Like law and medicine, data science is a vast and complicated field and much too big and complex for one person to master in one lifetime.

Similar to law and medicine, the emerging data science field contains a wide variety of skill sets. Like legal and medical specialties, each data science area can be divided into sub-specialties. Each data science specialty has differentiated knowledge and skill sets required to competently perform. Of course, certain knowledge and skill sets may be identical for multiple data science areas yet each specialty has a unique set of responsibilities and knowledge requirements that differentiate one specialty area from another. Swami Chandrasekaran created a visual representation of what it takes to become a data scientist (Figure 2).

Figure 2 Data science domains by Swami Chandrasekaran* (see online version for colours)



*Note that visualisation is used with permission by Swami Chandrasekaran

Mr. Chandrasekaran organised data science domains as follows:

- fundamentals
- statistics
- programming
- machine learning
- text mining/natural language processing
- data visualisation

- big data
- data ingestion
- data munging
- toolbox.

While Mr. Chandrasekaran represents each domain in a progressive fashion, it may be argued that – with rare exception – no one person can master each domain. In addition, I would add the following domains to Mr. Chandrasekaran's list:

- deep learning
- algorithm design and execution
- experimentation design and execution
- artificial intelligence
- probability theory – Bayesian probability
- forecasting techniques
- causation theory
- pattern recognition techniques
- data science power laws.

My Data Science Association colleague Gary Mazzaferro has been exploring the concepts and ideas surrounding data science and definitions as formalisations aligning with knowledge economies and the knowledge/science/technology maturity models. He has (to date) defined the following data science specialisations and types of data scientists:

- *Data Science*: A field of systematic interdisciplinary study to elucidate relationships across and within Formal, Social Natural and Special Sciences phenomenon through the application of scientific methods. Interdisciplinary areas include analytical processes, mathematics, probability and statistics, logic, modelling, machine learning, algorithms, communications, traditional sciences, business, public policy and philosophy.
- *Blue Sky Data Science*: A purely curiosity-driven exploratory branch of Data Science oriented towards the development and establishing understanding about relationships across and within phenomenon with no focus on specific goals and immediate application.
- *Basic Data Science*: A branch of Data Science research focused on clearly defined goals and oriented towards the development and establishing understanding about relationships across and within phenomenon.
- *Applied Data Science*: A branch of Data Science oriented towards the development of practical applications, technologies and other interventions including engineering practices. Applied Data Science bridges the gap between Basic Data Science and the

engineering domains to provide predictable, usable tools to industries including standard methods and practices.

- *Data Science Practice*: The regular performance of Applied Data Science activities and methods for private and public organisations. May practise externally or internally. Practice may necessitate additional disciplines based on the needs of the organisation including domain expertise and communications supporting presentation and reporting activities.
- *Data Scientist*: A person who studies or has expert knowledge of the interdisciplinary field of Data Science.
- *Blue Sky Data Scientist*: A person who studies or researches in the branch of Blue Sky Data Science.
- *Basic Data Scientist*: A person who studies, researches or has expert knowledge in the branch of Basic Data Science.
- *Applied Data Scientist*: A person who studies or researches in the branch of Applied Science.

Note that this is a preliminary list and not complete: we are at the very beginning of a long evolutionary process.

As a result, the evidence suggests the practice of data science will evolve into a profession where data scientists specialise in different areas – like lawyers and physicians. When you need to hire a lawyer, you usually consider the special area of law that a lawyer practises. If you have a tax problem, you hire a tax lawyer, not a divorce lawyer. If you have a heart problem, you do not hire a gynaecologist.

The profession of data science will evolve to create many specialisations. After all, it took law and medicine over 100 years to evolve as professions with different specialties.

5 How to organise data science as a profession

Different professions are organised differently. Considering that the authority to license and regulate causes a monopoly on rights, any profession may abuse authority in a self-serving manner under pretext of protecting the public. Law and medicine are examples of professions that both protect the public from unqualified practitioners and hurt the public in self-serving ways (e.g., restrict entry to raise pay and limit competition).

The goal is to design a data science profession that protects the public, advances data science as a science, improves the quality of data science services and creates an environment for rewarding careers. I respectfully suggest that we can learn from the legal and medical professions to borrow the good and mitigate the harmful.

Some consider the creation of professions under colour of law as motivated simply to create a monopoly to limit competition and raise prices. As a result, it is imperative for the emerging data science profession to show that self-regulation and setting performance standards improves the quality of services and protects the public notwithstanding secondary consequences of pay inflation or competition limits.

The following is a suggested series of seven process steps for the professionalisation of data science:

- the full-time occupation of data science is identified
- educational programs are established where specialised knowledge and skills are identified and incorporated into a data science curriculum
- a professional association is established to help define standards of the data science profession
- a professional code of ethics is created
- certifications and licences are developed to distinguish qualified from unqualified practitioners
- a professional association defines entry requirements and disciplinary procedures
- gaining the support of law for self-autonomy and self-regulation.

At this time, the full-time occupation of data science has been identified; educational programs are in the process of being established; the Data Science Association as professional association with over 4000 members has been established to help define professional standards and a data science professional code of ethics has been created.

Three important steps remain: data scientist certifications and licences; entry requirements and disciplinary procedures and support of law for self-autonomy and regulation. The time and difficulty in completing these remaining steps should not be underestimated.

At this time, the Data Science Association is developing educational guidelines and performance standards as well as a professional certification for data scientists. The data science professional certification is a process where a person proves that he or she has the knowledge, experience and skills to practise data science by passing an exam that is accredited by the Data Science Association.

The 'Data Science Professional Certification' demonstrates to employers and clients that a person is not only a qualified data scientist but also committed to data science as a profession. Certification makes a person more valuable to employers and allows data scientists to:

- enjoy better employment and advancement opportunities
- have a competitive advantage over candidates without certificates
- earn higher wages
- receive tuition reimbursement for continuing education.

The 'Data Science Code of Professional Conduct' of the Data Science Association provides ethical guidelines to help the data science practitioner. This demonstrates to employers and clients that a data scientist is not only committed to data science as a profession but will also act in the best interests of the client and prevent or mitigate potential harm to the client. It also protects data scientists from unscrupulous clients and employers who may desire to abuse data science to gain unwarranted or illegal advantage and potentially hurt the public.

In the future, the Data Science Association, in conjunction with other institutions in the data science community, will develop entry requirements, licensure and disciplinary procedures, and gain support of law for self-autonomy and regulation. This will take time

and hard work: it took law and medicine hundreds of years to become professions. We have the opportunity to make data science a profession in about ten (10) years.

All members of the data science community should collaborate in developing and professionalising data science to improve life, business and government. Together, we can shape a better future using data science!

Reference

Bullock, A. and Trombley, S. (1999) *The New Fontana Dictionary of Modern Thought*, Harper-Collins, London, p.689.

Bibliography

American Bar Association (2014) http://www.americanbar.org/about_the_aba/history.html

American Medical Association (2014) http://www.ama-assn.org/ama/pub/about-ama/our-history.page

Atlantic Magazine (2010) http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/2/

Brockman, N.C. (1962) *History of the American Bar Association*, 6 AM. J. LEGAL HIST. 3, 269–285.

Cox, L. (2010) *Creating a Profession and a Body of Knowledge for Product Supportability Engineering at High-Tech Companies*, Doctoral Dissertation, ABI-Inform 8 database.

Curnow, C. and McGonigle, T.P. (2006) 'The effects of government initiatives of occupations', *Human Resource Management Review*, Vol. 10, pp.286–292.

Data Science Association (2014) http://www.datascienceassn.org/code-of-conduct.html and http://www.datascienceassn.org/about-data-science

Ioannidis, J. (2005) *Why Most Published Research Findings Are False*, DOI: 10.1371/journal.pmed.0020124, http://www.datascienceassn.org/content/why-most-published-research-findings-are-false

Law, M. and Kim, S. (2005) 'Specialization and regulation: The rise of professionals and the emergence of occupational licensing regulation', *Journal of Economic History*, Vol. 65, pp.723–756.

McKinsey Global Institute (2013) Big Data: The Next Frontier for Competition, http://www.rosebt.com/uploads/8/1/8/1/8181762/big_data_next_frontier_for_innovation_competition_and_productivity.pdf

Merriam-Webster Dictionary (2014) http://www.learnersdictionary.com/definition/professionalize

Perks, R.W. (1993) *Accounting and Society*, Chapman & Hall, London.

United States Department of Labor, Bureau of Labor Statistics (2010) *Career Guide to Industries*, Retrieved on 3 February, 2012, http://www.bls.gov/oco/cg/cgs035.htm

Vollmer, H. and Mills, D. (Eds.) (1996) *Professionalization*, Prentice-Hall, Englewood Cliffs, NJ.

Waddington, I. (1990) 'The movement towards the professionalization of medicine', *British Medical Journal*, Vol. 301, No. 6754, pp.688–690.

Wikipedia (2014) http://en.wikipedia.org/wiki/Profession