Starting a Debate: Data Science – Occupation or Profession? A Discussion Paper.

Ursula Garczarek and Detlef Steuer

Abstract With this contribution at the ECDA-2019 in Bayreuth we want to start a much needed debate about the nature of the work of a data scientist. Is it a mere occupation or does the societal impact together with ethical issues surrounding the work imply data science should become a real profession in the sense of Airaksinen (Airaksinen, 2009). We explore the elements of data science and the responsibility a data scientist has for society. Some barriers are identified and what can be done about them. In this paper, we describe the line of reasoning which was presented, and some *lessons learned* from the actual discussions with the audience.

Ursula Garczarek
Cytel Inc, Clinical Research Services, International Chamber of Commerce (ICC)
Route de Pré-Bois, 20 C.P. 1839 , 1215 Geneva 15
Switzerland
✉ Ursula.Garczarek@cytel.com

Detlef Steuer
Helmut-Schmidt-Universität, Universität der Bundeswehr Hamburg
Holstenhofweg 85, 22043 Hamburg
Germany
✉ steuer@hsu-hh.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 6, No. 1, 2020

DOI: 10.5445/KSP/1000098011/16
ISSN 2363-9881

1 Starting the Debate

The talk at ECDA-2019 was given in a special session to stir up an open discussion with the audience. Here on one hand we summarise the talk, but also incorporate insights gained in the discussion.

The following statement shall be the starting point of a missing debate:

> Data science is in the focal point of current societal development. Without becoming a profession with professional ethics, data science will fail in building trust in its interaction with and its much needed contributions to society!

2 What Are We Talking About?

Before talking about the necessity of data science becoming a profession the important terms should be defined.

2.1 Data Science

We use the term data science in the sense of what Donoho (Donoho, 2017) calls *greater data science*. Donoho gives the following definition:

> *Data science is the science of learning from data. It studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner.*

The scope and impact of this science will expand enormously in coming decades as scientific data and data about science itself becomes ubiquitously available. (Greater) data science consists of

1. data gathering, preparation, and exploration,
2. data representation and transformation,

3. computing with data,
4. data modelling,
5. data visualization and presentation,
6. science about data science.

The last point is essential to define the scientific nature of a data scientist's work. From the discussions a big limitation of this definition was identified: It does not say anything about decisions to be made based on the work of data scientists. Given that the responsibility of data scientists is derived to a large extent from this influence on decision making, the definition of *data science* should include decision making.

A second lesson learned from the debate is, that starting with a definition splits rather than unites people involved in data science, because in different sub-groups – country, application area, academic vs industry etc – not only data science, but also the related areas of statistics, applied statistics, engineering, computer science, machine learning, do have different connotations.

Both limitations can be addressed in a first step by defining *data science tasks* rather than defining *data science*, e.g. by using the CRISP-DM framework (Shearer, 2000) as a start. What needs to be added then is the methodology development as being part of the *science* around the data science tasks, and the societal impact as being part of the *science of data science*. This is future work.

2.2 Profession

To understand the difference between occupation and profession we use the definition of (Airaksinen, 2009). A profession is different from an occupation in some characteristics:

- Scientific training
  Knowing what is to be done by understanding the rational, epistemological foundations of professional action.
- Autonomy
  A profession can influence the social decisions that regulate its members' work and their related rights and obligations.

- Professional ethics
  If the public needs the expertise of a group of specialists and therefore cannot unproblematically reject, challenge, or ignore the professional advice and the influence of their work, professional ethics becomes a key issue when the public evaluates the potential bias of professional work in relation to the quality of their life.

Especially the last point shows how data science must develop the self confidence to call itself a profession. The public in general can not challenge any conclusion a data scientist works out and, therefore, can not escape the consequences of that work. The influence on individual lives and society imply a big responsibility on the data scientist's side.

3 Has Data Science a Professional Nature?

In this section we show, that data science has all the features Airaksinen mentions to qualify as a profession.

Has Data Science a Societal Impact?

This is generally accepted. There is a vast amount of data collected about everybody's life. With today's computing power one can explore all that data very easily and cheaply! All that data is used for decision making on all aspects of human life.

Does Data Science Raise Ethical Issues?

An overview of the common ethical issues encountered using algorithms is given by the Commission Nationale de l'Informatique et des Libertés (CNIL, 2018). Six main challenges are enumerated in there we find very useful as a framework:

1. Autonomous machines: A threat to free will and responsibility.
2. Bias, discrimination and exclusion.
3. Algorithmic profiling: Personalising individual versus collective benefits.

4. Preventing massive files while enhancing AI: Seeking a new balance.
5. Quality, quantity, relevance: The challenges of data curated for AI.
6. Human identity before the challenge of artificial intelligence.

Not all issues have the same strength in connection to data science. Revisiting the definition of Donoho to relate the data science components to the issues in CNIL, one can see a lot of interconnections between these. Data science has a lot of influence on the design and the application of algorithms.

In the case of *autonomous machines*, the data scientist may not be involved in the final steps of building the machine. But without any doubt, there was a lot of data gathering, preparation and exploration before any part of the algorithm for autonomy of the machine was developed.

Bias, discrimination and exclusion are connected to the central theme of data science: The performance criteria used to build and assess algorithms. Similarly it's easy to find connections between any of Donoho's points and CNIL's challenges for safe applications of algorithms.

Influence (On Society) Leads to Responsibility

There are already examples of statistical and computer science societies emphasizing this responsibility. The point seems uncontroversial. To cite only a few interesting examples:

- ASA (American Statistical Society): Because society depends on informed judgments supported by statistical methods, all practitioners of statistics, – regardless of training and occupation or job title –, have an obligation to work in a professional, competent, respectful, and ethical manner (ASA, 2018).
- ACM (Association for Computing Machinery): Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good (ACM, 2018).
- GI (Gesellschaft für Informatik): GI members are especially committed to respecting and protecting human dignity. Whenever norms of the state,

> society or the private sphere come into conflict with these values, GI members must address the issue (GI, 2018).

So it seems data science by nature should be a profession. Why is it not recognized as such?

4 Principled Objections Against Data Science Ethics

If a discussion about ethical issues is started, often one is faced with very principled objections against formulating ethical rules or guidelines.

- Professional societies with memberships, codes of conducts asf. are elitist and limit scientific freedom and hinder innovation.
- What is right or wrong is defined by law.
- What is right or wrong is a personal matter.
- Corporate responsibility is more important than professional ethics, as those who use the implemented algorithms are responsible for their societal impact.
- Paper does not blush, and most code of conducts are hypocritical.

We like to call these objections the *lack of interest*. In the debate in Bayreuth, the ones on scientific freedom, and about corporate responsibility were shared by some, but seen rather as important considerations that need to be addressed (scientific freedom) or well-defined (professional responsibility vs. corporate responsibility) than that they would lead to a general objection to the development of data science ethics.

5 The Three Obstacles

In this section the three main obstacles in defining data science as a profession are discussed.

Lack of Knowledge

Today we find deficits in data science education mostly in the science on data science part. Airaksinen postulates a professional should be *knowing what is to be done by understanding the rational, epistemological foundations of professional action*. In practice, one sees data scientists with all sorts of basic education that are trained in three-month courses without any science on data science at all! For academically trained data scientists, – outside the few newly built data science study programs – the scientific embedding depends on their main subject, and there are at least two very distinct approaches to science on data science:

- The inferential framework in statistics, where data is seen as the result of some *data generating process* in the world, and the goal is to make reliable statements about that so-called world.
- The computational learning theory in machine learning, where data is seen as examples, and the goal is to learn a general concept from them that is optimal in a certain sense e.g. when applied with an algorithm to new examples.

There is no over-arching scientific framework for both of these or other relevant scientific embeddings of data science, though data scientists are approaching applications and using methodology from both frameworks. Furthermore, there is a lack of education on ethics. Only few study programs offer training on ethical reasoning or the moral or political aspects of data science.

At the conference this topic also was the subject of some controversial exchange, as the study programs for statistics, computer science, engineering, and the rather new data science programs, show a high degree of variation such that an oversimplification like here leads to protest from all those that are either completely overlooked or feel misrepresented. The learning is to focus on the goal of a more comprehensive scientific framework of data science, that includes data science ethics, and to see for any specific study program what is needed to reach it.

Lack of Communication Skills

At least statisticians are known for claiming to have excellent knowledge on how to do science on data science, but have limited success in communicating that knowledge back into society. There is a big need in society to understand algorithms and their consequences. Therefore data scientists need to learn to talk – and talk(!) – about methods and methodology with lay people!

Secondly, discussing ethics and our own moral views related to professional work should become a habit. To fulfill its responsibility to society, the data science community needs to overcome its shyness and/or ignorance to ethical considerations and debate, and make it an integral part of science on data science!

Lack of Power

At the moment individual data scientists lack power to successfully fight irresponsible data science in their day-to-day jobs. Due to a lack of defined standards, and the hype around data science, companies exist that sell expensive but less than useful data science services. Poor data science ruins the reputation of data science, even science in general, and ruins companies and institutions! But the position of those who try to fight against this is unnecessarily weak, when he or she is standing alone without a reference to generally accepted standards.

In companies or universities different sub-communities of data science exist in different departments, and often fight for funding. In the academic world there are still turf-battles (rather than an inspiring scientific debate) on being fundamentally right or wrong. To fulfill the requirements of a profession these different groups should join their powers!

6 Would *Professional* Data Science Help?

We are confident data science as a *profession* would help overcoming the obstacles just described.

Every individual data scientist would be supported by education and written guidelines on every day ethical decision making. She could refer to some manual instead of starting to think about the problems again and again.

Written rules of conduct for data science services would help to establish a relationship of trust between data scientists, their clients, their employers, and society.

Status, reputation and power of any individual data scientist would be increased, if data science acquired the trust of a profession. In the case of conflicts of interests an ethical guideline under the maintainership of some professional society may offer an arbitration process.

And not least it would be easier to fight back the charlatans, if the expectations and responsibilities of data science would be clearly defined and formulated. That is, if data science were a profession!

7 Conclusion: The Debate Is Much Needed!

Data science has all the features of a profession besides a professional ethic that would help even define data science itself. We hope to have clearly shown how such standards would help the individual, but also the community of data scientists as a whole, to build the trust society needs into our profession. The work has only just begun.

Postscriptum

The topic was considered important enough to be published, but potential reviewers were hesitant to accept the review as they felt they were no experts in the field. Thus now the paper is accepted without a formal review. We as authors shared the feeling of insecurity when writing, actually. This once more underlines the main point of the paper: It is important to have discussions about these topics inside our statistics and data science community to build the capability and the confidence that this is part of our profession! Otherwise statistics will be sidelined by computer science in shaping the future of data analysis.

References

ACM (2018) ACM Code of Ethics and Professional Conduct. URL: https://www.acm.org/code-of-ethics [accessed 2018-11-9].

Airaksinen T (2009) The Philosophy of Professional Ethics. In: Institutional Issues Involving Ethics and Justice, Elliot RC (ed), Encyclopedia of Life Support Systems (EOLSS), Vol. 1. Eolss Publishers, p. 201ff. ISBN: 978-1-848269-14-9. Developed under the Auspices of the UNESCO.

ASA (2018) Ethical Guidelines for Statistical Practice. URL: https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx [accessed 2018-11-9].

CNIL (2018) Algorithms and Artificial Intelligence: CNIL's Report on the Ethical Issues. Commission Nationale de l'Informatique et des Libertés. URL: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues [accessed 2018-11-2].

Donoho D (2017) 50 Years of Data Science. Journal of Computational and Graphical Statistics 26(4):745–766. DOI: 10.1080/10618600.2017.1384734.

GI (2018) Ethical Guidelines of the German Informatics Society. URL: https://gi.de/ethicalguidelines/ [accessed 2018-11-6].

Shearer C (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing 5:13–22. DOI: 10.4018/IJDWM.